Masther thesis

Assessing Differentiation in All Phases of Teaching: development of an assessment system for differentiated mathematics instruction in primary education.

Name student: Tjana Habermehl-Mulder, s1881736
To contact the student: a.t.mulder@student.utwente.nl
Name of university supervisor: Trynke Keuning / Marieke van Geel
To contact supervisor: t.keuning@utwente.nl / marieke.vangeel@utwente.nl
Keywords: differentiated instruction, assessment, development, primary education, mathematics
Word count: 12.411

Table of Content

Acknowledgements

In the first place, I want to thank certain people for their support during this research process. Regarding the support from the university:

T. Keuning – my first supervisor until she went on maternity leave. Thank you for the start of this process in which you guided me and for always telling me that it will turn out well.

M. van Geel – my first supervisor from the moment T. Keuning left. Thank you for the dedication with which you took over the guidance of me. Besides that, thank you for always challenging me to see things from all perspectives possible.

C. Smienk – who gave in me insight in the practical use of the ADAPT-instrument and in the theory behind it. Thank you for always willing to think along with me about the development of the instrument.

M. Dobbelaer – who helped me with questions about developing an assessment instrument and how to observe in the best way possible. Also, always wanted me to help with the statistical analysis software I did not understand in the first place.

For all four above, thank you for teaching me so much about differentiated instruction which will certainly help me in my own professional development as a teacher.

In addition, I would like to thank:

My husband, G. Habermehl – for always believing in me and supporting me in times I no longer thought I would make it. Thank you for all the times you helped me put things into perspective, so that I was confident again.

My parents – for their unconditional love and support throughout my entire school career.

Abstract

Within the MATCH-project, research about the concept of differentiated instruction (DI) in mathematics lessons in primary schools show that DI is a concept which occurs before, during and after the lesson and arises in the reasoning and acting of teachers (Van Geel et al., 2018). The ADAPT-instrument was developed to capture all phases into one assessment instrument. ADAPT stands for 'Assessing Differentiation in All Phases of Teaching' and entails an analysis of teacher documents, lesson observation and an interview tailored to the observed lesson. This study investigates whether such a complex instrument could meet validity and reliability criteria and still measures DI. During this study, the instrument was further developed in three phases with, as a result, recommendations for further development. In the first phase, the instrument was adjusted for scoring guidelines during a focus group of experts, followed by a training of raters. In the second phase, it was examined whether there was an inter-rater agreement. In addition, rater's comments were analysed together with some descriptive statistics. In the third phase the ADAPT-instrument was adjusted, during an expert meeting, based on the results of phase 2 and an improved version was developed. Then, this improved version was analysed based on the evaluation framework of Dobbelaer (2019) which can be used to evaluate the quality of the instrument and to evaluate the evidence gathered for reliability and validity of the instrument. The analysis confirmed that the inter-rater reliability should be tested again, when new raters are trained again for this improved version. Besides that, the interview of teachers about the observed lesson and rater manual need revision to cover all important issues of the instrument. Finally, the experts of phase 3 also mentioned future research should investigate if and how the mathematical domain of 'automation' should become part of the instrument. Overall, the ADAPT-instrument has made major quality steps towards reliability and validity criteria and, after further development, is expected to be able to assess the complex professional competency DI in all phases of teaching.

Keywords: differentiated instruction, assessment, development, primary education, mathematics

1. Introduction

In the MATCH-project, Keuning et al. (2017) and Van Geel et al. (2018) studied the complex competency 'differentiated instruction' (DI) in the field of mathematic lessons in primary education. They concluded that DI *during* the lesson cannot be separated from the phases of preparation and evaluation of the lesson. To capture DI as a whole, a cognitive task analysis (CTA) was conducted (Keuning et al., 2017; Van Geel et al., 2018) to distinguish the phases of DI and the teacher skills those phases entail. Based on the CTA, Keuning et al. (2017) and Van Geel et al. (2018) designed a professional development intervention, called the MATCH-project, to enhance the DI skills of (beginning) teachers during a mathematic lesson, based on all the phases of DI.

In addition, Van Geel et al. (2018) concluded that none of the operationalisations they have reviewed, captured the whole complexity of DI. Consequently, there was a need to develop a new adequate assessment instrument measuring the teachers quality of DI. A preliminary version (1.0) of the instrument for Assessing Differentiation in All Phases of Teaching (ADAPT-instrument) was developed. However, this version 1.0 is not tested yet for validation or reliability and the question arises whether this instrument meets these criteria to adequately assess DI or if it needs further development.

Assessment of professional competencies is very complex because a competency includes a complex integration of knowledge, skills and attitudes (Baartman, Bastiaens, Kirschner, & Vleuten, 2006). A problem that emerges, is to present evidence for validity and reliability for assessment instruments of complex competencies (Parsons et al., 2018), such as assessing all phases of DI. When an assessment instrument for DI is considered valid and reliable, it has the potential to be an important tool for many formative and summative purposes. To begin with, research by the Dutch Inspectorate of Education (Inspectie van het Onderwijs, 2018) showed a downward trend in mathematics results and underline the importance for teachers to develop their DI skills. As such, the ADAPT-instrument might be used as a formative feedback tool in professional development trajectories, like the MATCH-project.

On the other hand, the ADAPT-instrument could also entail a more summative purpose, for example, to monitor and evaluate the effectiveness of trainings, like the MATCH-project (Van Geel et al., 2018). Moreover, the purpose of the ADAPT-instrument could be to make high-stake summative evaluations by, for example, the Inspectorate of Education. Currently, the Dutch Inspectorate of Education (2018) assesses the quality of DI in primary schools very brief, with only four items and their ambition is to improve the educational results model of primary education (Inspectie van het Onderwijs, 2018). Furthermore, school leaders, school

boards or teacher education programs might want to use this instrument to assess, on low-stake summative basis, or to monitor their own (beginning) teachers.

This research aims to further develop version 1.0 of the ADAPT-instrument to make it more reliable and valid. In addition, the first exploration of the validity and reliability of the ADAPT-instrument will be encountered to evaluate the adjustments. When this instrument would be considered valid and reliable, it could serve one or more of the formative or summative purposes mentioned before.

2. Theoretical conceptual framework

In this section, first the concept of DI is discussed. Second, it will go deeper into what is already known about assessing complex professional competencies, such as DI. Then there will be made a connection between measuring DI and the development of assessment instruments with a view to reliability and validity, whereby reliability and validity will be explained and further investigated considering the assessment of DI.

2.1. Differentiated Instruction

Overall, it is hard to give a thorough description of DI, because it turns out to be a very complex teaching skill (Dixon, Yssel, McConnel, & Hardin, 2014; Eysink, Hulsbeek, & Gijlers, 2017; George, 2005; Grift, Wal, & Torenbeek, 2011; Keuning et al., 2017; Parsons et al., 2018; Van Geel et al., 2018). Parsons et al. (2018) described DI to be an ‘awesome balancing act’ in which “teachers adjust their teaching according to the social, linguistic, cultural, and instructional needs of their students” (p. 206). In general, DI is often described as the adaptions of aspects of instruction to differences between students (Bosker, 2005; George, 2005; Roy, Guay, & Valois, 2013). Nonetheless, Van Geel et al. (2018) reviewed thirteen operationalizations of DI and conclude these operationalizations:

> do not provide much insight into the acting and reasoning of teachers who differentiate instruction well. Such insight is required to measure differentiation as an aspect of teaching quality. In other words, we need to know what quality differentiation looks like as a basis for improving and assessing the quality of differentiation (Van Geel et al., 2018, pp. 3-4).

In short, in line with Parson et al. (2018), Van Geel et al. (2018) state that the reasoning and acting of a teacher are choices made before, during, and after instruction and therefore point out that DI is a concept which entails more phases than only DI during the lesson. To get insight into what those phases implies, Keuning et al. (2017) and Van Geel et al. (2018) performed a

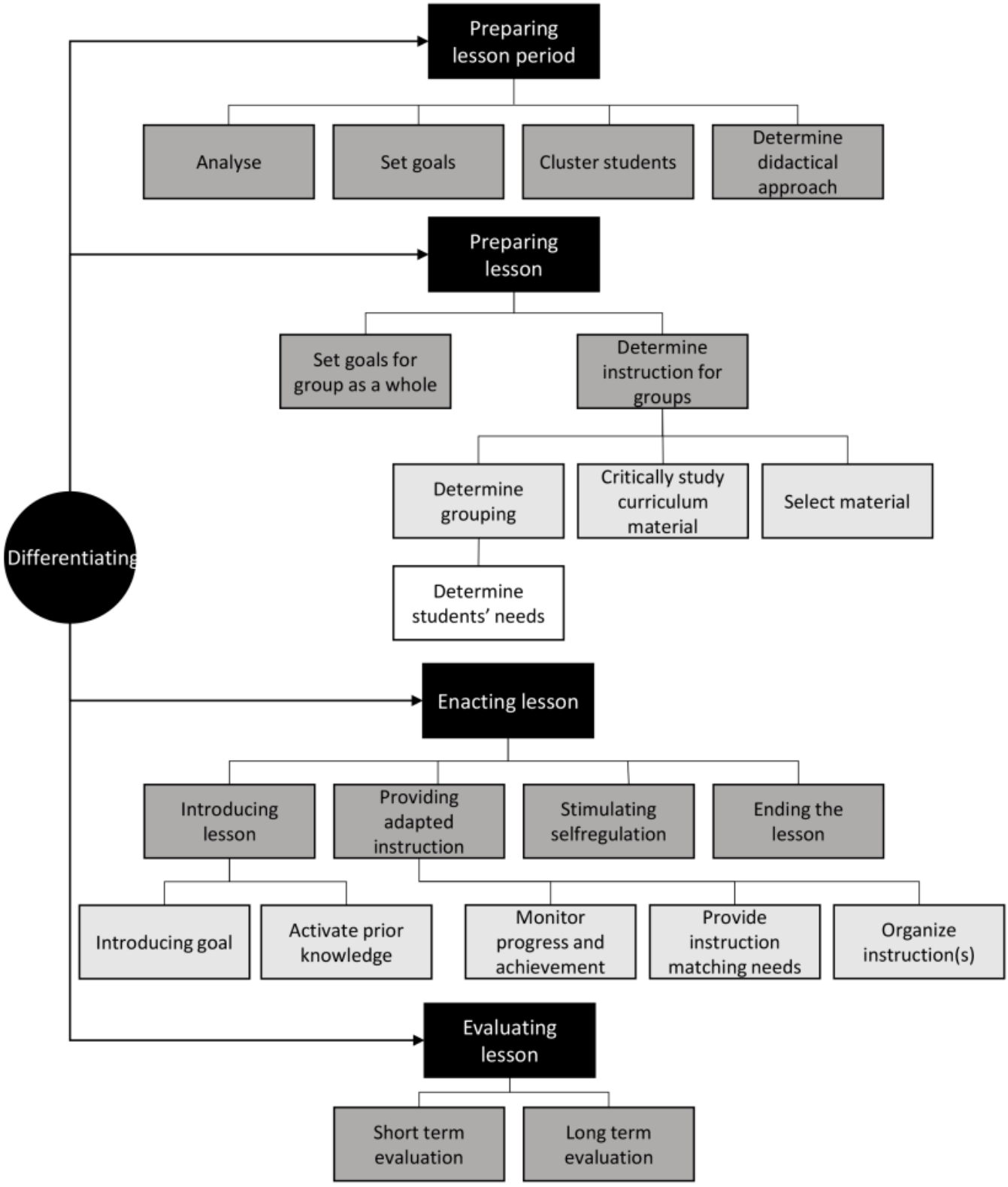


Figure 1. Differentiation skill hierarchy. Reprinted from "Capturing the complexity of differentiated instruction", by M. Van Geel et al., 2018, School Effectiveness and School Improvement, p. 10.

cognitive task analysis (CTA) focused on the actions and reasoning of teachers in all the phases of differentiation. Based on their CTA they distinguished the following four differentiation phases: *(1) preparation of the lesson period, (2) a teacher prepares a lesson, (3) teacher adequately address the differences between students during the lesson,* and *(4) the evaluation of the previous lesson*. Figure 1 depicts all four phases and shows that within each of these phases, several constituent differentiation skills can be distinguished (e.g. setting goals, determine instruction for groups, etc.). There is a temporal relationship among the horizontal adjacent constituent skills, "implying that they can be performed subsequently, simultaneously, or in a random order. Lower level skills facilitate the learning and performance of the skills higher up in the hierarchy" (Van Geel et al., 2018, p. 13). These findings prove DI occurs in more than one phase. Besides that, the results of the CTA showed that the key to successful differentiation is not to follow one specific kind of strategy but is in the deliberate and adequate choices a teacher makes, concerning the instruction. This assumes that when DI is assessed, those rationales should be taken in account (Van Geel et al., 2018). However, Van Geel et al. (2018) became aware of the fact that most of the reviewed operationalizations mainly consist of descriptions of applied differentiation strategies (e.g. grouping, varying assignments, etc.) and lack evaluation of the relationship between the instruction provided and the needs of students. The ADAPT-instrument is developed to capture all phases of DI and to incorporate the important relationship between observable actions and underlying rationales of teachers. Also, to evaluate the match between a teacher's actions and the student's needs. The question remains whether it is possible to assess such a complex skill in practice. The next section describes what is already known about assessing complex professional competencies, such as DI.

2.2 Assessing Complex Professional Competencies

To measure teacher quality in all phases of DI with all the related differentiation skills will certainly be very complex. Baartman et al. (2006) and Boudah (2011) declare, therefore, that assessing a professional competency, like DI, should entail more than one assessment method. Boudah (2011) calls this 'triangulation' which means the increase of truth value by using multiple assessment methods.

Besides that, Van Geel et al. (2018) added that assessment of DI would require "much time and effort from skilful assessor(s)" (p. 14). In their study to propose a model for designing assessment programs, Van der Vleuten et al. (2012) concluded that "we have no choice but to rely on the expert judgements of knowledgeable individuals at various points in the assessment process" (p. 207). They explain that expert judgement is needed to come to an aggregated

overall decision based on multiple (low-stake) assessments. For that reason, when multiple assessment methods would be used for assessing DI, it suggests that expert judgement is inevitable to come to an overall score.

When the interpretation of an observer plays an important role for scoring an instrument, adequate guidelines are needed to reduce subjectivity as much as possible (Dobbelaer, 2019). Therefore, Dobbelaer (2019) developed a framework which brings together the issues to be taken in account when developing, selecting, or using a classroom observation system (COS). In Figure 2, this framework is presented and shows the criteria for which instrument developers need to collect evidence. The framework is divided into three parts. The first part is meant for evaluating the characteristics of the COS, such as the theoretical basis, the quality of the instrument and the norms of the instrument. Each topic is divided into criteria for which evidence should be gathered when designing a COS. The second and third part of the evaluation framework are aimed at evaluating evidence for the reliable and valid use of a COS in a specific context. In the second part the focus lies on obtaining and reporting reliability evidence. The third part is about the evaluation of the validity argument in which Dobbelaer (2019) relates to the argument-based approach to validity (Kane, 2006; 2013). “In this approach, the network of inferences and assumptions leading from the sample of observations to the conclusions and decisions based on the observations is specified (the interpretive argument) and evaluated (the validity argument)” (Dobbelaer, 2019, p. 27). Considering the interpretive arguments, part three of the evaluation framework is divided in four common inferences: the scoring inference, the generalization inference, the extrapolation inference, and the implication inference. Each inference consists of warrants (rules and/or principles), divided in backing criteria (evidence to justify the warrant). Developers can use the framework to evaluate the quality of their instrument and to evaluate the evidence gathered for reliability and validity of the instrument (Dobbelaer, 2019). Dobbelaer (2019) underlines that it will be hard to meet all the indicators presented in the framework and therefore designers must decide “which evidence is most important for the reliable and valid use of the COS in their specific situation” (p. 33).

A few elements are important in the first developmental phase of an assessment instrument, according to Dobbelaer (2019). To generate observations that vary minimally between raters, developers should provide clear guidelines. First of all, the items of an instrument should not be unnecessarily difficult, to avoid scoring errors. This can be accomplished by adding “(1) scoring rules at the item level that help a rater distinguish between scores on a scale, and/or (2) scoring rules to compute an observed score (when multiple observations are conducted)” (Dobbelaer, 2019, p. 29). Dobbelaer states that to distinguish

which score to give and when, at the item level, can be achieved by adding scoring rubrics. An overall definition of the word ‘rubric’ is “a document that articulates the expectations for an assignment by listing the criteria or what counts, and describing levels of quality from excellent to poor” (Reddy & Andrade, 2010, p. 435). To compute an observed score, backing for scoring rules could be an observation protocol that is based on solid theory and /or rules that are supported by experts in the field (Dobbelaer, 2019).

Like mentioned before, the second and third part of the framework are about collecting reliability and validity evidence for that first step in the framework. As mentioned earlier, in the first developmental phase of an assessment instrument, scoring rules on the item level should be developed to ensure items are not unnecessarily difficult. Evidence for scoring rules on the item level can be gathered in the form of backing for the scoring inference. The scoring inference is about the claim that an observed score is generated based on a sample of observations (Dobbelaer, 2019). Gathered evidence for the scoring inference can be obtained by, for example, support of scoring rule(s) by experts (see warrant 1, evaluation framework) and/or rater training for scoring rules on the item level (see warrant 2, evaluation framework). Support of experts and/or rater training can be used as evidence for scoring rubrics and/or an observation protocol mentioned before. In line with Dobbelaer (2019), as mentioned earlier, the designer must decide what evidence is important to gather first in their specific situation.

Overall, to generate valid and reliable scores, developers have the primary responsibility in obtaining and reporting reliability and validity evidence (Dobbelaer, 2019). Therefore, in the next section the concepts of reliability and validity will be discussed and linked to the evaluating framework to investigate what additional evidence is needed for valid and reliable scoring rules.

Evaluation framework
Part A. Evaluation of the relevant COTAN criteria
1. Theoretical basis of the COS
1.1 Is the purpose of the COS specified?
a. Are the constructs that the COS intends to measure specified?
b. Is (are) the group(s) for which the COS is (are) intended specified?
c. Is the purpose of the COS specified?
1.2 Is the theory underlying the COS described?
1.3 Is the relevance of the COS's content for measuring the construct(s) justified?
2. Quality of the COS materials
2.1 Is the COS complete and clear?*
2.2 Are the items in the COS formulated correctly?
2.3 Is the scoring system devised in such a way that scoring errors can be avoided?*
3. Quality of the rater manual
3.1 Is a rater manual available?
3.2 Are the instructions for raters clear and complete?*
3.3 Is information provided on the applications and limitations of the COS?
3.4 Is a summary of the research findings published in the manual?
3.5 Is the degree of expertise required by raters to use the COS specified?
4. Norms
4.1 Are norms provided?
4.2 Are the norms up to date?
Norm-referenced interpretation
4.3 Are the norm groups large enough and representative?
Domain-referenced interpretation
4.4 Is there sufficient agreement between raters?
4.5 Have the raters been selected and trained appropriately?
5. Reliability
5.1 Is information on the reliability of the COS provided?
5.2 Are the findings of the reliability research sufficient considering the type of decisions that are intended to be made using the COS score?
5.3 What is the quality of the reliability research?
a. Are the procedures for computing the reliability coefficients correct?
b. Are the samples for computing the reliability coefficients consistent with the intended use of the COS?
c. Is the information provided sufficient to make a substantiated judgment of the reliability of the COS?
Part B. Evaluation of the validity argument
The scoring inference
Warrant 1: The scoring rule(s) is/are (statistically) substantiated
• The COS is based on theory, research, and/or standards
• The scoring rule(s) is/are supported by experts
• The scoring rule(s) is/are supported by teachers
• The scoring rule(s) has/have been tested in (pilot) research
• Statistical analyses support the scoring rule(s)
• The psychometric quality of the items is sufficient
Warrant 2: Measures were taken to score accurately and consistently
• For each criterion, scoring rules are available for raters
• Raters are trained in the use of the COS
• Raters are expected to meet a certain level of expertise
• Inter-rater reliability is sufficient
• Observation scores are consistent over time
Warrant 3: Attention is dedicated to preventing rater bias
• Attention is paid to preventing rater bias during rater training and/or in the COS manual
• Multiple raters are used to compute an observation score
• Statistical analyses show that raters do not rate specific groups of teachers differently from others
• Inter-rater reliability is sufficient
The generalization inference
Warrant: Opportunities for generalizations are explicitly described in the COS
• The required number of teacher observations is specified and substantiated
• The observation length is specified and substantiated
• The number of raters is specified and substantiated
• The observation moment is specified and substantiated

• The lesson type is specified and substantiated
• A generalizability study, a reliability study, or IRT analyses has shown that the sample of observations are representative of the assessment domain
• Research into the variation of observed lessons supports the generalizations of the observed score
• Confidence intervals are available and are based on a convention
The extrapolation inference
Warrant: The score in the assessment domain is related to the broader target domain
• The assessment domain covers a great portion of the target domain (face validity)
• The theoretical framework underlying the COS fits within the target domain
• The observed score is related to other measures within the target domain
The implication inference*
*Warrant 1: The implications for the observable attribute are appropriate (given the theoretical construct)**
*Warrant 2: The implications for the observable attribute are appropriate (given the statistical analyses)**
Note: * data not used in analyses

Figure 2. Evaluation Framework. Reprinted from "The quality and qualities of classroom observation systems", by M. J. Dobbelaer, 2019, Doctoral dissertation, University of Twente, p. 146-148.

2.3. Reliability

Reliability is the degree to which a study can be exactly repeated by independent researchers (Boudah, 2011; Carmines & Zeller, 1979; Hernon & Schwartz, 2009; Vos, 2009). Boudah (2011) divides this concept into *internal reliability* (the degree to which data collection, analysis, and interpretations are consistent under the same conditions) and *external reliability* (the extent to which an independent researcher could replicate the study in other settings). Boudah (2011) explains it is crucial to identify ‘’the reliability of the measure chosen for evaluating the dependent variable in a study’’ (p. 71) before investigating the reliability of the study as a whole. In other words, in this case it is important that the assessment instrument must be consistent under the same conditions (internal reliability), before analysing the reliability of the instrument when it would be used by independent researchers (external reliability). Internal reliability can be distinguished in two major areas: reliability of an instrument and reliability of observation (Boudah, 2011). To measure the reliability of an instrument, a reliability coefficient is needed to indicate the relationship between multiple items, administrations, or other analyses of evaluation measures. Reliability of observations can be calculated by an inter- or intraobserver agreement and by an inter- or intrascorer agreement. The question is which issue of reliability should be identified first when developing an assessment instrument. Dobbelaer (2019) states it is inherently relevant to gain information about the inter-rater reliability when

an observation instrument is used by raters. Also, Reddy and Andrade (2010) suggest, when scoring rubrics are added, using rater reliability to show if a rubric lead to a relatively common interpretation. When this is not the case, the items within the scoring rubric need revision and/or the rater needs training. Dobbelaer (2019), Van Vleuten (2016) and Vos (2009) underline the importance of the latter in which, to reduce errors, rater training would eliminate individual differences in the way raters decide which score to give.

This indicates that after defining scoring rules, such as designing a scoring rubric and / or good training of raters, the first step is to measure the inter-rater reliability, indicating whether each rater would give the same score under the same conditions. Besides that, it will give crucial information about which items of the instrument need revision.

Nonetheless, when an inter-rater agreement would be achieved, this would not mean the instrument truly measures the quality of DI. When an item is considered relatively reliable, it is not also relatively valid (Carmines & Zeller, 1979; Kirk & Miller, 1986; Vos, 2009).

2.4. Validity

Like reliability, Boudah (2011) and Hernon and Schwartz (2009) divide validity into *internal validity* (does the instrument measure what it intends to measure) and *external validity* (can the findings be generalized to a larger population). In the evaluation framework of Dobbelaer (2019) the scoring inference (internal validity) is analysed before the generalization inference (external validity) and therefore the focus for this study lies on the internal validity. Also, to develop a new questionnaire, Trochim and Donnelly's (2006) state first the construct of an instrument needs to be valid. Figure 3 shows their framework for construct validity, in which a construct must fulfil both translation and criterion-related validity requirements. Translation validity involves content validity (whether the constructs are theoretically well defined and inclusive) and face validity (which focuses on the clarity of items, based on the theoretical constructs). Criterion validity entails a more relational approach in which the construct proved the conclusions that are expected, based on the theory. Criterion validity involves convergent validity (items of a construct should be highly correlated to each other) opposed to discriminant validity (in which items from different constructs should not be highly correlated to each other). Criterion validity involves also predictive validity (the construct should predict something it should theoretically predicts) and is opposed to concurrent validity (a construct should distinguish between groups when it should theoretically be able to distinguish). First it should be ensured that an instrument is well founded in theory (content & face validity) before ensuring the relational approach (criterion validity).

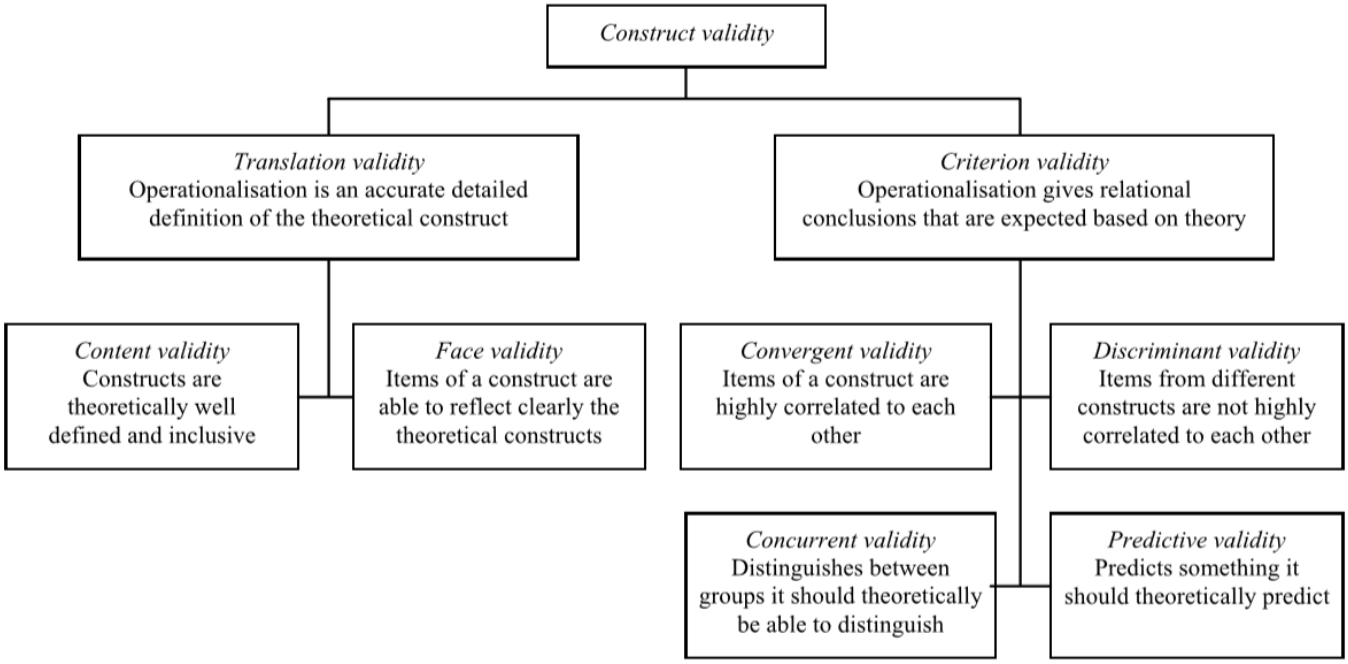


Figure 3. Framework for construct validity. Reprinted from *The research methods knowledge base,* by Trochim, W. M., & Donnelly, J. P., 2006, Cincinnati, OH: Atomic Dog.

Comparing this framework of construct validity of Trochim and Donnelly (2006) to the evaluation framework of Dobbelaer (2019) a certain similarity can be seen in which step to take first when designing an assessment instrument. Dobbelaer (2019) mentions the constructs need to be specified and founded in theory (part one of the framework), which is in line with the part of translation validity (content & face validity) of Trochim and Donnelly (2006). In part three of the evaluation framework of Dobbelaer (2019) evidence for the validity argument needs to be gathered and starts with the evaluation of the scoring inference. Scoring inference ‘’connects a sample of observation to an observed score’’ (p. 29). To build that argument, Dobbelaer mentions that the scoring rules should be appropriate, and supported by relevant stakeholders like experts in the field. Experts in the field can examine whether a construct and / or a corresponding item covers the essential topic (content & face validity) (Vos, 2009). This indicates that when an assessment instrument is developed, first experts should examine the instrument to maximise content and face validity (translation validity).

Besides that, it stands out that inter-rater reliability, as described earlier for reliability, is also mentioned in the third part for validity for scoring inference, in the evaluation framework of Dobbelaer (2019). She states that when the inter-rater reliability is sufficient it would be backing for the scoring inference in this way that the instrument can be used accurately in a specific context and that raters can use the instrument consistently over time.

3. Research question and model

This research will analyse version 1.0 of the ADAPT-instrument, for assessing teachers' quality of differentiated mathematics instruction in primary schools, and further develop this instrument, based on theory mentioned earlier. This further developed ADAPT-instrument, version 2.0, will be tested in a pilot research. Results of that pilot research will be used to further develop the instrument into a version 3.0. This version 3.0 will be analysed, according to the evaluation framework of Dobbelaer (2019), to give recommendations for further development. This leads to the following research question and sub questions:

3.1. Research question

How can an assessment instrument of the complex competency 'teachers' quality of differentiated mathematics instruction in primary schools' be further developed?

3.2. Sub questions

- What is the inter-rater reliability of the ADAPT-instrument, when version 1.0 is further developed by adding scoring rules in the form of scoring rubrics?
- Which issues of the ADAPT-instrument and /or scoring guidelines need adjustments, based on results of first raters as well as expert judgement regarding content and face validity?
- To what extent does the new developed version 3.0 of the ADAPT-instrument meet the evaluation framework of Dobbelaer (2019)?

3.3. Scientific and practical relevance

Keuning et al. (2017) and Van Geel et al. (2018) captured the concept of DI as showed in Figure 1. However, no prior operationalizations of DI have examined the acting and reasoning of teachers during all phases (before, during, and after instruction) (Van Geel et al., 2018). Van Geel et al. state:

> 'given the complexity of differentiating in itself and the interrelatedness of a variety of aspects involved in quality differentiation, the question remains whether and, if so, how we can assess this complexity in an efficient manner within the reality of the school context'' (p. 13).

This study will contribute to answering this question, by evaluating and improving reliability and validity of version 1.0 of the ADAPT-instrument, and by identifying barriers to assess DI.

As stated earlier, the ADAPT-instrument can serve formative and summative purposes, when proven valid and reliable. Dijkstra et al. (2012) state to choose the purpose in the following way: ''the higher the stakes, the more robust the information needs to be'' (p. 5).

When in the future, the ADAPT-instrument is valid and reliable in all ways possible, high stake summative decisions can be made as mentioned in the introduction. However, when evidence for reliability and validity is not as robust as necessary for summative decisions, the instrument can still be able to serve formative low-stake decisions, like formative feedback in professional development trajectories.

4. Research design and methods

4.1. Research design

Several steps will be taken to refine version 1.0 of the ADAPT instrument and to obtain evidence for reliability and validity. Those steps can be distinguished into three phases. In the first phase scoring rubrics will be added to the ADAPT-instrument and will result in version 2.0, the second phase entails testing the reliability and usability of this version 2.0 of the ADAPT-instrument. In the third phase, an improved version 3.0 of the ADAPT-instrument will be created based on the outcomes of phase 2. That version 3.0 will be analysed according to the evaluation framework of Dobbelaer (2019) for recommendations of future research. The different phases are explained further in the procedure.

This study is a mix-method of quantitative and qualitative descriptive research in which a condition is described (Boudah, 2011). Quantitative, because it attempts to describe empirically whether the further developed ADAPT-instrument, version 2.0, will lead to a high degree of inter-rater reliability. The content and face validity will be measured in a qualitative way (joint expert/research meeting) as such it includes data which is retrieved through observation, interview, and document review (Boudah, 2011). Both the quantitative and qualitative data will be used to develop and refine version 3.0 of the ADAPT-instrument.

4.2. Procedure

Phase 1. First, the researcher will make a ‘rubric set-up’ for the ADAPT-instrument with a performance level descriptor per item, based on a consultation with the trainer in the MATCH-project who used version 1.0 before. In addition, the researcher of this study and four experts in the field (two researchers and one trainer of the MATCH-project and one expert in development of assessment instruments) will develop version 2.0 of the ADAPT-instrument within a focus group.

Phase 2. Phase two focuses on testing the inter-rater reliability and usability of version 2.0 of the ADAPT-instrument developed in the first phase. First, the focus group of phase 1 will train together with the raters of this phase for assessing the quality of DI of teachers with

this version 2.0 of the ADAPT-instrument. For the training, data will be used from teachers who are not included in the sample of this study.

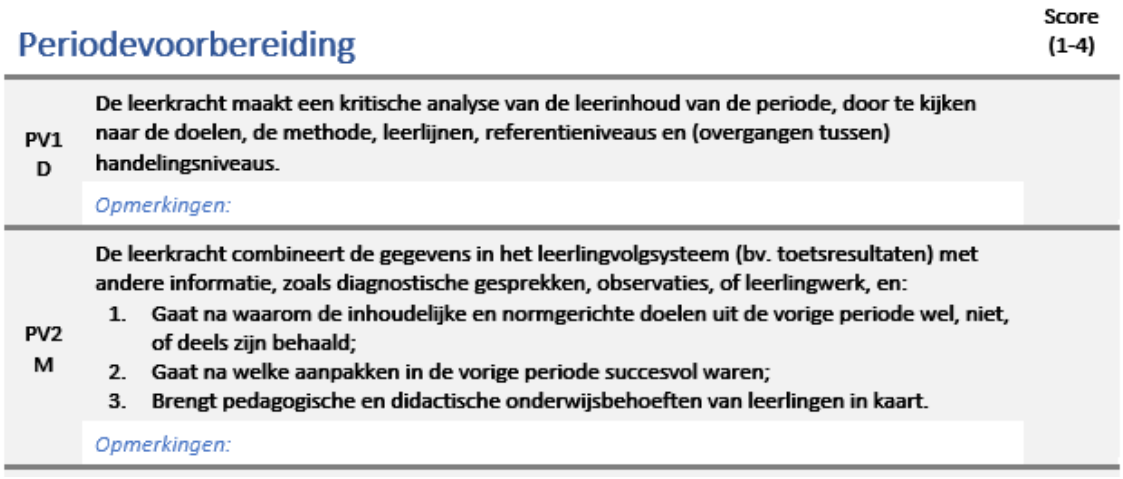
Next, for the inter-rater reliability mathematics lessons of 17 teachers will be scored in random order. The data of the teachers includes two video-taped mathematics lessons, one interview tailored to the first lesson and additional data (optional) like period plans and/or instructional plans. Raters have access to the password-protected datafiles, however datafiles cannot be downloaded due to privacy of the teachers. The researcher will score all 17 teachers and two other raters, from the focus group in phase 1, will score 9 and 8 teachers respectively. Raters will register their scores in an online scoring form in which the rater can also provide comments per item to explain the given score. Also, an additional question is added to provide other comments about the instrument itself and its usability. Findings of phase 2 will be used for the development of a version 3.0 of the ADAPT-instrument in phase 3.

Phase 3. In this phase version 2.0 of the ADAPT-instrument will be refined into a version 3.0. Adjustments will be made based on the results of phase 2 and the evaluation framework of Dobbelaer (2019), within a focus group of the researcher and two experts of phase 1. In addition, recommendations will be given for further development based on the outcomes of phase 3 and the evaluation framework of Dobbelaer (2019).

4.3. Instruments

Version 1.0 of the ADAPT-instrument will be used in phase 1 for adjustments mentioned in the procedure. This version of the ADAPT-instrument (see Appendix A) was developed in and for the MATCH-project. The content of the instrument was derived from performance objectives, which are based on the CTA performed by Keuning et al. (2017) and Van Geel et al. (2018). The instrument consists of a combination of observation and interview. An interview guide is included to collect information to score all items. In particular, to score the items for other components than what can be observed in the lesson itself.

The first page of the instrument serves as rater manual. It explains what the instrument and the abbreviations entail. In addition, it describes that as score of 1 to 4 must be given and that it is up to the rater to use the knowledge of the circumstances and characteristics of the teacher's work situation to link a sound value judgment to the teacher's performance. Therefore, it is mentioned that comments are necessary to understand the reasoning of raters in giving a teacher a certain score.



Periodevoorbereiding		Score (1-4)
PV1 D	De leerkracht maakt een kritische analyse van de leerinhoud van de periode, door te kijken naar de doelen, de methode, leerlijnen, referentieniveaus en (overgangen tussen) handelingsniveaus. *Opmerkingen:*	
PV2 M	De leerkracht combineert de gegevens in het leerlingvolgsysteem (bv. toetsresultaten) met andere informatie, zoals diagnostische gesprekken, observaties, of leerlingwerk, en: 1. Gaat na waarom de inhoudelijke en normgerichte doelen uit de vorige periode wel, niet, of deels zijn behaald; 2. Gaat na welke aanpakken in de vorige periode succesvol waren; 3. Brengt pedagogische en didactische onderwijsbehoeften van leerlingen in kaart. *Opmerkingen:*	

Figure 4. Design of version 1.0 of the ADAPT-instrument. Ech row represents one item, in this case from the phase 'preparation of the lesson period'.

The rest of the instrument consist of items for all the phases of DI: 8 items for *preparation of the lesson period* (indicated with 'PV'); 7 items for *a teacher prepares a lesson* (indicated with 'LV'); 10 items for *teacher adequately address the differences between students during the lesson* (indicated with 'L(number of lesson)LU'), divided into the introduction, core and end of a lesson; and 2 items for *the evaluation of the previous lesson* (indicated with 'EV'). Figure 4 depicts the first two items of version 1.0 for the phase *a teacher prepares a lesson*. The abbreviations that stand before each item, represents the phase: PV1 stands for item 1 of 'Preparation of the lesson period' (*Periodevoorbereiding*). The letter underneath the abbreviation, e.g. the 'D' underneath 'PV1', represents a corresponding differentiation principle, based on the differentiation skill hierarchy (see Figure 1) from Van Geel et al. (2018). The letters represent the following principles: 'goal-oriented (D) *(doelgericht werken)'*, 'challenging (U) *(uitdagen)'*, 'monitoring (M) *(monitoren)'*, 'adjusting (A) *(afstemmen)'*, and 'promotion of self-regulation (Z) *(zelfregulatie stimuleren)'*. On the right, a score must be given per item on a scale of 1 till 4 in which score 1 means 'point of attention' and score 4 means 'excellent'. The rater can, in addition, explain the score given in the space for comments *(opmerkingen*). Besides, the ADAPT-instrument has an appendix with 'explanatory notes (*toelichtingen criteria*)', also derived from the performance objectives which are based on the CTA of Keuning et al. (2017) and Van Geel et al. (2018). The explanatory notes consist of examples in practice per item and can be consulted by raters (see Appendix B). Figure 5 depicts the explanatory notes of PV1 and PV2, where the abbreviations have the same meaning as in the instrument and mentioned before.

Toelichtingen criteria periodevoorbereiding

PV1 D	**De leerkracht maakt een kritische analyse van de leerinhoud van de periode, door te kijken naar de doelen, de methode, leerlijnen, referentieniveaus en (overgangen tussen) handelingsniveaus.** De leerkracht vormt een beeld van wat er in de komende periode aan bod komt en hoe dit door de methode/software wordt overgebracht. De leerkracht legt verbanden tussen de doelen die aan bod komen en algehele leerlijn. Er wordt nagegaan of er doelen worden overgeslagen, welke doelen nieuw zijn, welke cruciaal zijn en welke nog vaker terug gaan komen.
PV2 M	**De leerkracht combineert de gegevens in het leerlingvolgsysteem (bv. toetsresultaten) met andere informatie, zoals diagnostische gesprekken, observaties, of leerlingwerk, en:** **1. Gaat na waarom de inhoudelijke en normgerichte doelen uit de vorige periode wel, niet, of deels zijn behaald;** **2. Gaat na welke aanpakken in de vorige periode succesvol waren;** **3. Brengt pedagogische en didactische onderwijsbehoeften van leerlingen in kaart.** Zowel methodegebonden als methode-onafhankelijke toetsresultaten moeten hierbij worden betrokken (evt. een voortoets). Een eerste grove analyse kan gebeuren door bijvoorbeeld een 3x3 matrix te maken, die kan worden aangevuld met een specifiekere analyse op basis van de categorieënanalyse, vaardigheidsscore, niveauscores en vaardigheidsgroei. Indien doelen niet behaald zijn wordt er actief naar een reden of oorzaak gezocht. Tevens worden in het overzicht van onderwijsbehoeften stimulerende en belemmerende factoren genoteerd.

Figure 5. Design of version 1.0 of the *explanatory notes*, representing two items from the phase 'preparation of the lesson period'.

Finally, there are guideline questions which the interviewer of the MATCH-project used during the interview with the teachers (see Appendix C).

In phase 2 the further developed version of the ADAPT-instrument, based on results of phase 1 will be used. In short, the results of every phase, are the instruments of the next phase and for that reason not yet possible to describe.

4.4. Participants

Phase 1. As mentioned in the procedure, in this phase there will be made use of the expertise of four people of which three are involved in the MATCH-project and one is the developer of the evaluation framework of Dobbelaer (2019). All are female, of which all four have a master's degree focused on education. In addition, three graduated from a teacher training academy (PABO) for primary education. One expert is nowadays trainer of the MATCH-project. Two others are postdoctoral researchers at University of Twente and currently researchers in the MATCH-project. The last expert is the developer of the evaluation framework of Dobbelaer (2019) and expert in developing assessment instruments.

Phase 2.

Raters. In this study three raters scored teachers with the ADAPT-instrument. All raters (females) graduated from a teacher training academy (PABO) for primary education. The first rater (age = 32 / 0 years of teacher work experience), has a masters' degree in Educational

science and works at the University Twente as a doctoral candidate. The second rater (age = 34 / 3 years of teacher work experience) has a masters' degree and PhD in Educational science and works at the University Twente as a postdoctoral researcher. The third rater (age = 24 / 1.5 years of teacher work experience) is a master student of *Educational Science and Technology* at the University Twente.

Sample. The data of 17 teachers was retrieved, with permission, from a project of the MATCH-project, from teachers of two primary schools in the province of Overijssel (11 teachers) and Gelderland (6 teachers) in the Netherlands. The average age of the teachers was 44 years (M = 43.53; SD = 15.03) ranging from 26 to 64 years. The average years of work experience was 21 (M = 21.29; SD = 15.30). Table 1 shows descriptive statistics of the participants.

Table 1

Sample description.

	n	%	*M* (*SD*)
Gender			
Male	3	17.65	
Female	14	82.35	
Age			43.53 (15.03)
Years of work experience			20.29 (15.30)
Educational level (%)			
HBO (PABO)	16	94.1	
postHBO	3	17.6	
Academic university	0	0.00	

Phase 3. In this phase the outcomes of phase 2 will be used to develop an improved version of the ADAPT-instrument together with the expertise of two experts of the MATCH-project from phase 1. The two experts are one postdoctoral researcher of the University Twente and the trainer of the MATCH-project.

4.5. Data analysis

Phase 1. The researchers' set up of the rubrics was analysed and, at the same time, adjusted in the expert meeting.

Phase 2. To estimate the inter-rater reliability Dobbelaer (2019) advices to use the statistical test Cohen's Kappa. This test measures the percentage of agreement between raters, adjusted for agreement by chance (Vos, 2009; Dobbelaer, 2019). For the development of the ADAPT-instrument inter-rater reliability on the item level will be analysed with Cohen's Kappa. When two raters appears to disagree often on one item, it probably means this item needs revision. These items will, consequently, be discussed in the focus group of phase 3. Besides that, some descriptive statistics will be used to gain further insight in the differences between raters in their usage of the ADAPT-instrument. Also, the comments given per item will be reviewed and analysed. This is a qualitative method in order to determine patterns or themes (Boudah, 2011).

Phase 3. During this meeting of experts, version 2.0 of the ADAPT-instrument will be adjusted to a version 3.0 based on outcomes of phase 2. In addition, the researcher will analyse version 3.0 of the ADAPT-instrument according to the evaluation framework of Dobbelaer (2019) to give recommendations for the next step of development.

5. Results

5.1. Results Phase 1

Together with the trainer of the MATCH-project, the researcher made a set-up of a rubric for the ADAPT-instrument and the *explanatory notes*. The agreement was to make a performance level descriptor per item and include them into the instrument. Every separate performance level descriptor represented a score of 1, 2, 3, or 4. The content of each performance level descriptor was based on the item, as formulated in version 1.0 of the ADAPT-instrument and based on the performance objectives obtained from results of the CTA (Keuning et al., 2017; Van Geel et al., 2018) and the experience of the MATCH-trainer. In order to see what distinguished the content of each performance level descriptor from the previous performance level descriptor, the 'new' content was given a different colour, and this was explained and described at the first page of the instrument, which serves as rater manual. Next, a step-by-step-plan was added which served as an observation protocol for each rater in how to assess each teacher and to make sure each rater assesses in the same way for reliability purpose. As a result, there was a first concept of the second version of the ADAPT-instrument, which can be described as concept 1.1 of the ADAPT-instrument.

Subsequently, concept 1.1 of the ADAPT-instrument and the *explanatory notes* were analysed and adjusted in a focus group consisting of the researcher and four experts mentioned in the participants section. In three days, each item and its matching performance level descriptors were analysed, adjusted and finetuned into version 2.0 of the ADAPT-instrument and version 2.0 of the *explanatory notes* (see Appendix D and E). Two of the experts were the researchers and developers of the differentiation skill hierarchy of Van Geel et al. (2018) and their expertise was used to make sure the content still reflects DI as a whole. The other expert was the researcher and developer of the evaluation framework of Dobbelaer (2019) and her expertise was used to give recommendations to make the instrument as valid and reliable as possible.

In Figure 6, the first two items of version 2.0 of the ADAPT-instrument are presented, where the abbreviations still mean the same as in version 1.0. Scores can be given of 1 till 4, where the blue colour shows the difference between the current performance level descriptor and the previous one, e.g. the blue text in performance level 3 descriptor is added to the performance level 2 descriptor. The scores stand for *poor (1), insufficient (2), sufficient (3), good (4),* respectively. However, there was made a remark that sometimes information is lacking to give a deliberate score. For example, information is lacking when during the interview a certain aspect of DI is not discussed. In that case there is also the option to score 'could not be assessed (*niet te beoordelen*)'. Besides that, sometimes the item is not applicable, e.g. when there are no students with a higher level of math, consequently, there are no goals for them. In that case the option 'not applicable (*niet van toepassing*)' can be chosen. *Could not be assessed* and *not applicable* are not an option for every item, because most items can always be assessed or are always applicable.

Periodevoorbereiding

		1	2	3	4	Score	Niet te beoordelen
PV1 M	**De leerkracht evalueert of de doelen van de vorige periode zijn behaald en waarom (niet). Hiervoor gaat de leerkracht na welke aanpakken in de vorige periode (niet) succesvol waren.**	De leerkracht bekijkt alleen behaalde scores (bijv. niveauscore op LOVS-scores en/of cijfers op methode toetsen).	De leerkracht bekijkt de behaalde scores. De leerkracht bepaalt op basis hiervan of de doelen wel, niet of deels zijn behaald.	De leerkracht bekijkt de behaalde scores. En relateert de scores aan de inhoudelijke doelen. De leerkracht bepaalt op basis hiervan of de doelen ook inhoudelijk wel, niet of deels zijn behaald.	De leerkracht bekijkt de behaalde scores. En relateert de scores aan de inhoudelijke doelen. De leerkracht bepaalt op basis hiervan of de doelen ook inhoudelijk wel, niet of deels zijn behaald. De leerkracht gaat na waarom de doelen uit de vorige periode wel, niet, of deels zijn behaald. En gaat hierbij na welke aanpakken in de vorige periode succesvol waren.		
	Opmerkingen:						
PV 2 M	**Brengt pedagogische en didactische onderwijsbehoeften van leerlingen uitgebreid in kaart.**	De leerkracht brengt de onderwijsbehoeft-en van leerlingen niet in kaart.	De leerkracht brengt alleen algemene pedagogische of didactische behoeften van leerlingen in kaart.	De leerkracht brengt alleen algemene pedagogische en didactische behoeften van leerlingen in kaart.	De leerkracht heeft een breed beeld (bijv. op basis van leerlingwerk, diagnostische gesprekken en dagelijkse observaties) van pedagogische en didactische behoeften van de leerlingen.	Score	n.t.b.
	Opmerkingen:						

Figure 6. Design of version 2.0 of the ADAPT-instrument. Ech row represents one item, in this case from the phase ‘preparation of the lesson period’.

Toelichtingen criteria periodevoorbereiding

PV1 M	**De leerkracht evalueert of de doelen van de vorige periode zijn behaald en waarom (niet). Hiervoor gaat de leerkracht na welke aanpakken in de vorige periode (niet) succesvol waren.**	
	1	Hierbij worden alleen scores geconstateerd.
	2	Je kijkt hierbij vooral normgericht. Ze moeten bijv. 80% norm gehaald hebben.
	3	Hier wordt ook inhoudelijk naar de doelen gekeken. Een categorieanalyse, wat zowel op LOVS & methodetoetsen kan.
	4	Hierbij wordt ook een verklaring gegeven voor het wel of niet behalen van de norm/inhoudelijke doelen. Een leerkracht constateert dat de plusleerlingen een bepaald doel niet hebben behaald, omdat hij/zij ze tijdens deze lessen niet aan de instructie heeft laten meedoen.
PV2	**Brengt pedagogische en didactische onderwijsbehoeften van leerlingen uitgebreid in kaart.**	
	1.	n.v.t.
	2	Bijv. dat een leerling baat heeft bij het krijgen van stapsgewijze instructie (algemene behoeftes). Het gaat hier om dat het alleen of pedagogische **of** didactische behoeftes zijn.
	3	Het gaat hier om dat het pedagogische **en** didactische behoeftes zijn.
	4	Een breed beeld krijgt de leerkracht door bijv. een diagnostisch leergesprek te voeren om in kaart te brengen welke voorkeursstrategie een leerling heeft om cijferend te vermenigvuldigen.

Figure 7. Design of version 2.0 of the *explanatory notes*, representing two items from the phase ‘preparation of the lesson period’.

In addition, the *explanatory notes* were adjusted, in which for each item a performance level descriptor with an example in practice was included (see Figure 7). According to the experts, when no example was necessary, the content of the performance level descriptor is 'n.v.t. (*not applicable)'*. The *explanatory notes* were for consultation of raters and not mandatory to use.

After finalizing version 2.0 of the ADAPT-instrument, the focus group trained together for calibration of the use of the instrument. The instrument has been completed twice, for two teachers. Next, phase 2 started with gathering data for exploring the inter-rater reliability. In addition to this instrument, a scoring form was developed as aid for raters to have a clear overview of the scores they gave (see Appendix F).

5.2. Results Phase 2

Inter-rater reliability. In total there were three raters: TH, the researcher; MD, the doctoral candidate; MG, the postdoctoral researcher. TH assessed all 17 teachers, MD assessed 9 teachers, and MG assessed 8 teachers. For the 9 teachers MD and TH scored, two mathematics lessons were observed and assessed. For the 8 teachers MD and TH scored, only the first lesson of each teacher was observed and assessed. Whereas rater MD and TH assessed 6 teachers in total and for the last 2 teachers only the items of the phase *teacher adequately address the differences between students during the lesson.* Appendix G contains the raw scores, per item, from raters MD-TH and MG-TH for the teachers and the corresponding comments that the raters gave.

The inter-rater reliability was calculated for the rater pairs MD-TH and MG-TH with Cohen's Kappa. Table 2 shows the results of the Kappa's for raters MD and TH per item divided in different Kappa results and the same accounts for Table 3 with the results of raters MG and TH. Each column with 'K_....' represents a Kappa result and, where Kappa calculates the agreement between raters, adjusted for the element of chance (Vos, 2009). A 0 implies that agreement is equivalent to chance and 1 stands for perfect agreement. For the results in Table 2 and 3, .41 - .60 reflects moderate agreement (*k > .41) and .61 – 1 reflects substantial to perfect agreement (**k > .61).

However, sometimes a rater had no variance in scores for an item and therefore kappa is considered to be zero. No variance means that one rater on one item has given all teachers the same score. This does not necessarily mean that there was no agreement between raters and for that reason the 'Raw Agreement' was calculated in those cases. When calculating the raw agreement, the number of agreements in scores is divided by the total (*n).* Same as with kappa, 0 means no agreement and 1 means perfect agreement, only Raw Agreement is not adjusted for

chance. For the results in Table 2 and 3, .41 - .60 reflects moderate agreement (*RA > .41) and .61 – 1 reflects substantial to perfect agreement (**RA > .61).

Since different answer options were possible, this must be taken into account in the analysis. Therefore, different kappa's are calculated for different circumstances per item. First, the kappa for all scores was calculated (K_Overall). This K_Overall shows to what extent the raters agreed to give a score of 1, 2, 3, 4, *could not be assessed* or *not applicable*. Subsequently, a logical sequence has been followed in analysing the inter-rater reliability where secondly, kappa was calculated for all times raters agreed that an item could be assessed or not (K_Can be assessed). Third, when raters agreed that a score could be assessed, kappa was calculated for all times raters agreed about whether the item is applicable or not (K_Applicable). Fourth, when an item could be assessed and is applicable, kappa was calculated for all times raters agreed on the score to give (K_Judgement). Because there is a difference in scores between scoring insufficient (score 1-2) and sufficient (score 3-4) for an item, kappa was also calculated for the cases raters agreed about this difference (K_(In)sufficient).

Raters MD and TH. When analysing kappa's and Raw Agreements of raters MD and TH the following stands out. Overall, there is an agreement about whether an item could be assessed or not (27 times k or RA > .61 and 4 times k or RA > .41) and total substantial agreement about whether an item is applicable or not (37 times k or RA > .61). However, agreement in exact score is low. Only two items (LV6 & L2LU5) stand out by having a kappa (k) or Raw Agreement (RA) > .61 for K_Judgement. The items PV6, L1LU3, L1LU9 and L2LU6 scored k or RA >.41 for K_Judgement. However, a k or RA of .41 - .60 out of 9 times (or less, because K_Judgement = n – cases with *could not be assessed (5) – not applicable (6)*) suggests those items need revision. Moreover, comparing this to the total of 37 items raters MD and TH scored, only 2 items have a high k or RA which suggests that the other 35 items and/or corresponding performance level descriptors need revision. Only, 10 items of the phase *teacher adequately address the differences between students during the lesson* are scored twice by raters MD and TH because they assessed two mathematics lessons of each teachers. This means that in total 25 items are suggested that need revision.

Regarding the kappa's and Raw Agreement of K_(In)sufficient, both raters agreed 11 times with k or RA >.61 and 5 times with k or RA >.41. 11 times substantial agreement out of 37 is less than half of times and shows that raters MD and TH often disagree whether a teacher scored (in)sufficient for an item.

Raters MG and TH. In Table 3, the results of raters MG and TH are presented. These raters scored one mathematics lesson per teacher and so 27 items are scored in total. Overall, also

rater MG and TH agreed most of time whether items could be assessed or not (18 times *k* or *RA* > .61 & 2 times *k* or *RA* > .41) and total substantial agreement till perfect agreement about whether an item is applicable or not (27 times *k* or *RA* > .61).

For K_Judgement, both raters only scored *k* or *RA* > .61 for one item (PV6) and for items PV2, PV8, L1LU1, LILU6, L1LU7, L1LU8, and EV2 they scored *k* or *RA* > .41. However, as mentioned before, only one item with *k* or *RA* > .61 could be considered 'good', which implies that all other 26 items need revision.

When analysing K_(In)sufficient, also raters MG and TH agreed more often, compared to K_Judgement. Of 9 items *k* or *RA* was >.61 and 9 items *k* or *RA* was >.41. 9 times substantial agreement out of 27 items is, in line with results of raters MD-TH, less than half and shows that raters MG and TH often disagree whether a teacher scores (in)sufficient for an item.

Across raters. In conclusion, the output of results of the two pair of raters, MD-TH and MG-TH, are generally the same. Both did not agree often for K_Judgement and agreed more for K_(In)sufficient, however this applied to less than half of the items. It would be interesting to investigate further, by analysing the comments, why raters differ in their judgements. Both scored, for K_Judgement, a relatively high *k* or *RA* for items PV6 (*k* or *RA* > .61) and LU6 (either lesson 1 or 2; *k* or *RA* > .41), which suggest that those items do not need revision. In addition, a first cautious conclusion could be that all other 25 items need revision in phase 3. To get better insight in what causes so much difference in scores, the variance in scores between raters will be investigated next.

Table 2

Inter-rater reliability of raters MD and TH

	Item	*n*	K_ Overall	RA	*n*	K_Can be assessed	RA	*n*	K_Appli-cable	RA	*n*	K_ Jugde-ment	RA	*n*	K_(In) suffi-cient	RA
PV1	Evaluatie	9	.26		9	.77**		5	.00	1**	5	-.18		5	.00	.40
PV2	Behoeften in kaart brengen	9	.24		9	.40		3	.00	1**	3	-.20		3	.00	.67**
PV3	Kritische analyse	9	.30		9	.31		4	.00	1**	4	.33		4	.20	
PV4	Reparatie-doelen	9	.21		9	1**		6	1**		6	-.13		6	-.33	
PV5	Verrijkings- en/of verdiepings-doelen	9	.27		9	.77**		5	.00	1**	5	-.11		5	-.36	
PV6	Clustering van leerlingen	9	.63**		9	.73**		6	.00	1**	6	.60*		6	.57*	

PV7	Organisatori-sche en didactische aanpak	9	.16	9	.50*	5	.00	1**	5	-.11	5	.29		
PV8	Zelfregulatie	9	.36	9	.57*	4	.00	1**	4	.20	4	.00	.75**	
LV1	Lesdoelen formuleren	9	.16	9	.73**	6	.00	1**	6	.00	.00	6	.00	.14
LV2	Instructie-groepen	9	.24	9	.73**	6	.00	1**	6	.00	6	.25		
LV3	Sterke rekenaars	9	.30	9	.73**	6	.00	1**	6	.11	6	.18		
LV4	Passende instructie	9	.40	9	.77**	5	.00	1**	5	.17	5	.62**		
LV5	Passende verwerking	9	.50*	9	.78**	4	.00	1**	4	.27	4	.50*		
LV6	Zelfregula-tie van leerlingen	9	.84**	9	1**	6	1**		6	.71**	6	.00	.83**	

LV7	Les mentaal doorlopen voor hulpvragen	9	.25		9	.57*		4	.00	1**	4	.00	.25	4	.00	.50*
L1 LU1	Introductie lesdoel	9	.23		9	1**		9	1**		9	.23		9	.00	.89**
L1 LU2	Voorkennis activeren	9	.26		9	1**		9	1**		9	.26		9	.25	
L1 LU3	Kwaliteit basis-instructie	9	.0	.44*	9	1**		9	1**		9	.00	.44*	9	.00	.56*
L1 LU4	Monitoren	9	.12		9	.00	.89**	9	.00	1**	8	.13		8	.25	
L1 LU5	Onver-wachte gebeurte-nis(sen)	9	.22		9	.53*		2	.00	1**	2	.00	.00	2	.00	1**
L1 LU6	Verlengde instructie	9	.43*		9	.73**		6	1**		5	.00	.40	5	.00	.40

L1 LU7	Sterke rekenaars	9	-.03	9	-.17	6	.00	1**	6	.00	6	.08	
L1 LU8	Balans instructie & verwerking	9	1	9	1**	9	1**		9	.10	9	.36	
L1 LU9	Zelfregula-tie van leerlingen	9	.51*	9	1**	9	1**		9	.51*	9	1**	
L1 LU 10	Werkproces & lesdoel evalueren	9	.33	9	1**	9	1**		9	.33	9	.25	
EV1	Lesdoelen evalueren (korte termijn)	9	.29	9	.53*	5	.00	1**	5	.17	5	.29	
EV2	Reflectie leerkracht (lang termijn)	9	.37	9	1**	3	1**		3	-.29	3	.0.	.33

L2 LU1	Introductie lesdoel	9	.17	9	1**		9	1**		9	.17	9	1**	
L2 LU2	Voorkennis activeren	9	.24	9	.00	.78**		.00	1**	7	.34	7	.42	
L2 LU3	Kwaliteit basis-instructie	9	.18	9	1**		9	1**		9	.18	9	.25	
L2 LU4	Monitoren	9	-.08	9	1**		9	1**		9	-.75	9	.00	.78**
L2 LU5	Onver-wachte gebeurte-nis(sen)	9	1**	9	1**		0	1**		0	1**	0	1**	
L2 LU6	Verlengde instructie	9	.03	9	-.20		5	.00	.80**	4	.00	.50*	4	.00	.50*
L2 LU7	Sterke rekenaars	9	-.29	9	-.24		2	.00	1**	2	.00	.00	2	.00	.00
L2 LU8	Balans instructie & verwerking	9	-.11	9	1**		9	1**		9	-.11	9	.05	

L2 LU9	Zelfregula-tie van leerlingen	9	.36	9	1**		9	1**		9	.36	9	.73**
L2 LU 10	Werkproces & lesdoel evalueren	9	.29	9	.00	0,89 **	8	.00	1**	8	.33	8	.20

Note. Abbreviations in item row represent the phase, e.g. PV1 stands for item 1 of 'Periodevoorbereiding' *(preparation of the lesson period);* RA = Raw Agreement; K_Overall = Cohen's Kappa of the overall scores; K_Can be assessed = Cohen's Kappa if raters agree whether the item could be assessed or not; K_Applicable= Cohen's Kappa if raters agree whether the items (than can be assessed) are applicable or not; K_Judgement = Cohen's Kappa if raters agree in the scores that are applicable; K_(In)sufficient = Cohen's Kappa if raters agree that items are sufficient or insufficient.
k = **k* > .41. ***k* > .61.
RA = * *RA* > . 41. ***RA* > .61

Table 3

Inter-rater reliability of raters MG and TH

	Item	*n*	K_ Overall	RA	*n*	K_Can be assessed	RA	*n*	K_Appli-cable	RA	*n*	K_ Jugde-ment	RA	*n*	K_(In) suffi-cient	RA
PV1	Evaluatie	6	.00		6	.33		2	.00	1**	2	.00	.00	2	.00	.00
PV2	Behoeften in kaart brengen	6	-.14		6	.00		2	.00	1**	2	.00	.50*	2	.00	1**
PV3	Kritische analyse	6	-.03		6	1**		6	1**		6	-.03		6	.00	.50*
PV4	Reparatie-doelen	6	-.11		6	.00		2	.00	1**	2	.00	.00	2	.00	.00
PV5	Verrijkings- en/of verdiepings-doelen	6	.25		6	.57*		4	.00	1**	4	-.33		4	.00	1**
PV6	Clustering van leerlingen	6	.20		6	.00	.50*	3	.00	1**	3	.00	.67**	3	.00	.67**

PV7	Organisatori-sche en didactische aanpak	6	-.03		6	-.20		4	.00	.75**	3	-.20		3	.00	.67**
PV8	Zelfregulatie	6	.20		6	-.29		3	.00	.67**	2	.00	.50*	2	.00	1**
LV1	Lesdoelen formuleren	6	.00	.17	6	1**		6	1**		6	.17		6	.00	.83**
LV2	Instructie-groepen	6	-.13		6	1**		6	1**		6	-.13		6	-.20	
LV3	Sterke rekenaars	6	.06		6	1**		6	1**		6	.06		6	.08	
LV4	Passende instructie	6	.10		6	.00	.83**	5	.00	1**	5	.23		5	.55*	
LV5	Passende verwerking	6	.25		6	.00	.83**	5	.00	1**	5	.33		5	.55*	
LV6	Zelfregula-tie van leerlingen	6	.10		6	.00	.83**	5	.00	1**	5	.17		5	.00	1**

LV7	Les mentaal doorlopen voor hulpvragen	6	.09		6	.33		3	.00	1**	3	.00	.00	3	.00	.00
L1 LU1	Introductie lesdoel	8	.48*		8	1**		8	1**		8	.48*		8	.71**	
L1 LU2	Voorkennis activeren	8	.27		8	1**		8	1**		8	.26		8	.60*	
L1 LU3	Kwaliteit basis-instructie	8	.24		8	1**		8	1**		8	.24		8	.60*	
L1 LU4	Monitoren	8	.22		8	1**		8	1**		8	.22		8	.25	
L1 LU5	Onver-wachte gebeurte-nis(sen)	8	.00	.63**	8	1**		6	.00	.83**	0	-		0	-	
L1 LU6	Verlengde instructie	8	.48*	8	8	1**		8	1**		8	.56*		8	.71**	

L1 LU7	Sterke rekenaars	8	.00	.00	8	1**	8	.00	.88**	7	.00	.00	7	.00	.14
L1 LU8	Balans instructie & verwerking	8	.56*		8	1**	8	1**		8	.56*		8	.60*	
L1 LU9	Zelfregula-tie van leerlingen	8	.37		8	1**	8	1**		8	.37		8	.60*	
L1 LU 10	Werkproces & lesdoel evalueren	8	.33		8	1**	8	1**		8	.33		8	.00	.50*
EV1	Lesdoelen evalueren (korte termijn)	6	.33		6	1**	6	1**		6	.33		6	.33	
EV2	Reflectie leerkracht (lang termijn)	6	.36		6	.33	3	.00	1**	3	.50*		3	1**	

Note. Abbreviations in item row represent the phase, e.g. PV1 stands for item 1 of 'Periodevoorbereiding' *(preparation of the lesson period);* RA = Raw Agreement; K_Overall = Cohen's Kappa of the overall scores; K_Can be assessed = Cohen's Kappa if raters agree whether the item could be assessed or not; K_Applicable= Cohen's Kappa if raters agree whether the items (than can be assessed) are applicable or not; K_Judgement = Cohen's Kappa if raters agree in the scores that are applicable; K_(In)sufficient = Cohen's Kappa if raters agree that items are sufficient or insufficient.
k = **k* > .41. ***k* > .61.
RA = * *RA* > . 41. ***RA* > .61

Variance in scores. After investigating the agreement across raters, variance in scores was explored. First, the cases where *could not be assessed* or *not applicable* was scored were excluded from the dataset. Only the cases where a score was given remained and so *n* was different for each item. Next, the variance in scores between raters was statistical analysed and displayed in bar charts for each item (see Appendix H for all bar charts).

Figure 7 displays the two bar charts of variance in scores for L1LU7. The bar charts of L1LU7 reflect the other results and illustrates, to a certain degree the explanations of the results from the kappa's and Raw Agreements. It illustrates that most of the time there is a difference of 1 or 2 between scores. A difference of -1 / +1 indicates that raters gave adjacent scores, and -2, -3 / +2, +3 indicates that scores are far apart. The difference of -1 / +1 can be either sit within the judgement (in)sufficient (score 1 and 2, or score 3 and 4), as outside it (score 2 and 3). Therefore, colours are added in the bar charts whereby the red colour means that raters did not agree about the teacher scored (in)sufficient for that item (score 2 and 3) and blue means raters did agree about the teacher scored (in)sufficient for that item (score 1 and 2 or score 3 and 4). What stands out is that the distribution of the colours is spread and there is no structure in it. This confirms the kappa and Raw Agreement results of K_(In)sufficient where for both pair of raters, MD-TH and MG-TH, the agreement was less than half of the time.

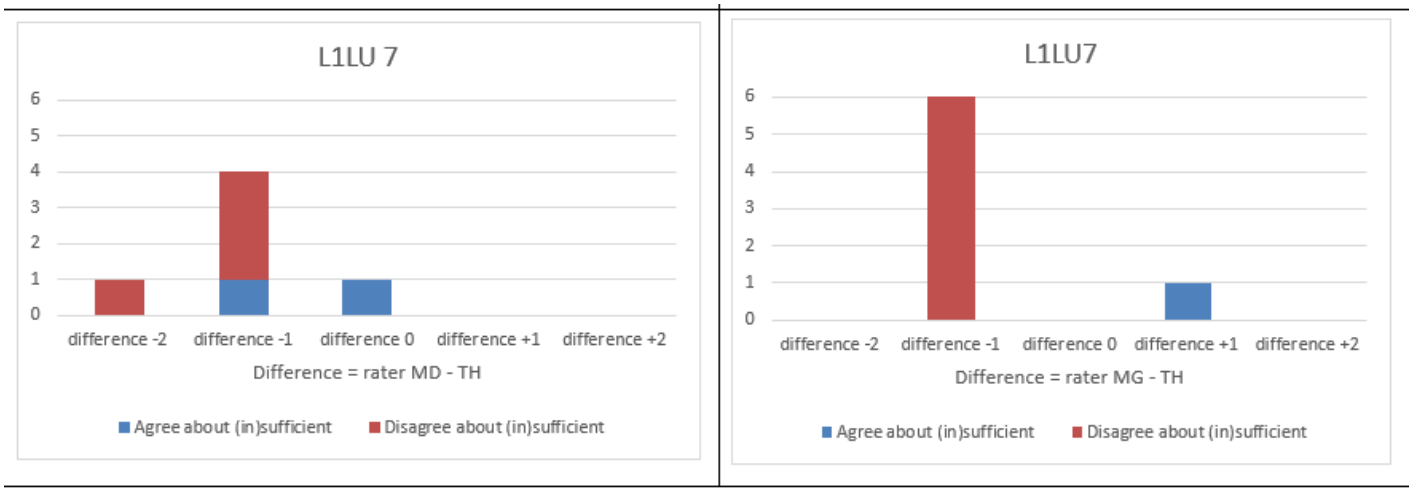


Figure 8. Variance in scores of item L1LU7 of raters MD-TH and MG-TH.

Another result that stands out is that rater TH structurally gave higher scores relative to raters MD and MG. This becomes apparent because most variances are negative (e.g. difference -2 or difference -1). The fact that the researcher TH gave higher scores cannot be explained by the variance but might be explained by the comments. Possible reasons for giving higher scores

can be that rater TH is the researcher of this study and therefore fully dedicated, that she is a teacher herself, that she is younger, that she is less experienced in performing scientific research/data gathering, that she is less critical and/or that the performance level descriptors are not clear enough and this result is a coincidence. Bringing this in line with earlier results, what can be concluded is that when all the items that need revision in phase 3 are revised, the variance in scores should be less than it is now in phase 2. On the other hand, a low *n* and difference in scores can also be due to a lack in information and will be discussed next.

Could not be assessed. When there was not enough information to give an item a score, raters filled in '*could not be assessed'*. Table 4 shows how many times this score was filled in per rater, in total and per item. The table is divided in two timeframes. At the beginning of the year, 9 teachers were observed and were assessed by rater TH and MD. In the middle of the year the other 8 teachers were observed and assessed by rater TH and MG.

It appears that in phase *preparation of the lesson period* (PV) the frequency of score *'could not be assessed'* is structurally high. This assumes that most information was missing in this phase. This lack of information can be due to the fact that it was not discussed in the interview, teacher documents were lacking, and/or for the phase *teacher adequately address the differences between students during the lesson* the quality of the filmed lesson was not good enough to see certain elements needed to score the 'PV' items.

For the phase *a teacher prepares a lesson* only for the teachers which MD and TH rated there was often a lack of information. The reason why this occurred less for the teachers rated by MG and TH is probably because these teachers had a different interviewer, who asked more about this topic. Another possible explanation is that these interviews took place at a later stage in the year than the interviews with the first 9 teachers. This could also explain why the percentages '*could not be assessed*' by MG-TH are lower in the other phases.

The higher percentages in the first two and the last phases (PV, LV and EV) suggests that information from the interview is necessary. Besides that, for the phase of *teacher adequately address the differences between students during the lesson* (L(number of lesson)LU) results also show that for some items information was lacking. This can be explained in the way that it is not seen on the video-tape but could also mean that information from the interview (which was not available now) is crucial to score these items.

Now it is known that inter-reliability for most of the items it not achieved and that scores vary between raters in most of the times with a score of +1/-1. Also, a lot of information was lacking during this phase of scoring. The comments of raters are not taken in account yet and might give crucial insight into what were the underlying intentions when giving a certain score.

Table 4

Percentage of the score 'could not be assessed' per rater

	9 teachers – beginning of the year						8 teachers- middle of the year					
	Rater						Rater					
	TH			MD			TH			MG		
Items	*n*	CBA	%	*n*	CBA	%	*n*	CBA	%	*n*	CBA	%
Total	333	90	27.3	333	80	24.0	216	24	11.1	182	16	8.8
PV	72	30	41.7	72	25	34.7	64	15	23.4	48	12	25
LV	63	23	36.5	63	22	34.9	56	4	7.1	42	3	7.1
L1LU	90	8	8.9	90	13	14.4	80	2	2.5	80	0	0
EV	18	10	55.6	18	8	44.4	16	3	18.8	12	1	8.3
L2LU	90	20	22.2	90	12	13.3						

Note. CBA = cannot be assessed; PV = preparation of the lesson period (*periodevoorbereiding)*; LV = teacher prepares a lesson (lesvoorbereiding); L1LU = first lesson of phase: teacher adequately address the differences between students during the lesson (*lesuitvoering)*; EV = the evaluation of the previous lesson (evaluatie); L2LU = second lesson of phase: teacher adequately address the differences between students during the lesson (*lesuitvoering).*

Rater Comments. Appendix G contains the raw scores and all the comments per item. All comments are analysed by the researcher, and some stand out in relation to the scores given and will be discussed next. First comments per rater pair will be analysed separately to see, in the end, what kind of results are the same for all raters.

MD and TH. When analysing the comments of raters MD and TH, a few things stand out. Table 5 shows the raw scores for item LV4 about the preparation of the instruction and findings in this item are representive for findings in the other items.

Comments belonging to teacher A and B show, like mentioned before, that without a proper interview, this item cannot be scored and for that reason *cannot be assessed* is filled in.

Another issue concerning the score *cannot be assessed*, for example to be seen at teacher H, is that it is not clear when to give this score. Where MD gives a score of 2 for the information given, TH decided information was insufficient to give a score. This happened 37 times for rater pair MD-TH and suggests that in phase 3 it would be wise to define a scoring rule for when to give this score of *cannot be assessed.*

For teacher D and E, scores differ with a score of 1 but both raters agreed about being sufficient. This difference in score is probably on one hand due to subjectivity of the rater and on the other hand due to ambiguity of the items and/or the performance level descriptors. An example of this is teacher E where both raters give the same comment but score different. The question remains whether the items and/or performance level descriptors are ambiguous or if the raters unintentionally depended more on their own opinion instead of following the performance level descriptors strictly.

For teacher I, both raters clearly interpreted information different, looking at the scores and comments. Rater MD is of opinion that this teacher barely prepares the lesson and rater TH lists a few things teacher I mentioned in the interview and therefore gives the highest score, according to the performance level descriptor. This issue also addresses the problem of subjectivity and/or ambiguity of items and/or performance level descriptors mentioned before. On the other hand, it might be that the items and/or the performance level descriptors are clear, but the information from the interview can be interpreted differently. To avoid this problem as much as possible, questions of the interview should be close to the content of the performance level descriptors.

In the same way, however, it is important that raters analyse all information possible, otherwise it can mistakenly seem that information is lacking. An example can be seen at item PV3 for teacher H, where rater MD mentions there is no mention of pedagogical needs of students in the interview. However, rater TH mentions the teacher document where this is

described for the students and therefore TH gives a higher score. This shows that it is crucial to make connections between given data to give a fair score and this might be an issue for training raters in the future and/or could be better reflected in the rater manual.

Table 5

LV4 – Appropriate Instruction

Teacher	Rater MD	Rater TH	Comments MD	Comments TH
A	CBA	CBA	Ik weet hier niks over.	Ze bereidt de instructie een dag van te voren voor en is er is een vast format voor de les. Maar echt inhoudelijk over hoe de instructie is voorbereid is niet gesproken
B	CBA	CBA	Geen interview beschikbaar	Geen nagesprek aanwezig
C	2	2	volgt methode	Kort, tijd en materiaal wordt bekeken.
D	3	4	Bekijkt hoe het het beste aangeobden kan worden	Deel afgeweken van de methode. Past naar eigen inzicht aan. Methode is middel, niet doel. Echt handelingsniveau wissel & methode wissel.
E	4	3	Kijkt vtv even naar de les en welke concrete materialen nodig zijn (ook voor sterkere rekenaars)	Onderzoekt het en zorgt dat handelingsniveaus kloppen door bijv. materiaal erbij te pakken
F	CBA	CBA	Geen interview beschikbaar	Geen nagesprek gedeeld.
G	2	2	Houdt voornamelijk vast aan de methode en probeert dat nog wat meer betekenisvol te maken of een koppeling met concrete materialen	Neemt het altijd over van methode. Voegt misschiennog wat materiaal toe & context, maar de methode volgt ze vast. Kan wel ook tijd aanpassen.
H	2	CBA	Ze kijkt alleen wel/geen opwarmer. Verder volgt ze voornamelijk de methode	Niet goed genoeg over gehad.
I	1	4	Ze bereid het nauwelijks voor	Handelingsniveaus verschillend toepassen. Instr. verkort. Zwakke lln extra instr. & sterke verkort.

Note. Abbreviation CBA stands for '*cannot be assessed*'.

MG and TH. Table 6 shows the results of rater MG and TH for item PV4 about setting 'repairing' goals for students and is, overall, representive for the other results. What stands out first is that rater MG did not give a lot of comments. This makes it complicated to analyse the differences in scores.

Nonetheless, there are results worth mentioning like the results for teacher J, L, N, and P. When rater TH did not have enough information to give a fair score she scored *cannot be assess,* while rater MG scored for example a 1. The other way around is seen for teacher N where rater TH gives a score and rater MG states there is not enough information. This discrepancy occurred 26 times for rater pair MG-TH. As previously stated, scoring rules are needed for raters to know when a score of *cannot be assessed* should can be given. Additionally, results of teacher M shows again that items and/or performance level descriptors need to be less ambiguous. Where rater MG says she doubts about scoring a 2 or 3, rater TH decides to give a score 3. However, this could also be due to unclear information, causing doubt with rater MG.

Additionally, for the phase *teacher adequately address the differences between students during the lesson,* rater MD and TH did sometimes give the comments like 'see PV…' and/or 'see LV...' *(zie PV …/zie LV …).* This finding suggests that information of preparation phases is needed to give a score in the phase of *teacher adequately address the differences between students during the lesson*.

Table 6

PV4 – Repairing Goals

Teacher	Rater MG	Rater TH	Comments MG	Comments TH
J	1	CBA	x	Kan niet precies zeggen of en vooral hoe ze dit doet.
K	2	3	x	x
L	CBA	3	x	Omdat het zo goed in de blokvoorbereiding naar voren komt wie waar nog moeite mee heeft (bijv. vanuit het vorige blok)
M	2	3	Twijfel 2/3	Elke dag instr van de vorige dag herhalen voor zwakke lln.
N	3	CBA	x	OPP sommigen lln, maar of per periode echt herhalingsdoelen voor 1-ster lln zijn gemaakt, weet ik niet precies.
O	CBA	CBA	x	ZE hebben snappet, maar hoe de doelen tot stand komen, voor zwakke lln is niet duidelijk genoeg besproken.
P	1	CBA	x	Kan niet precies zeggen of en vooral hoe ze dit doet.
Q	2	3	x	x

Note. Abbreviation CBA stands for '*cannot be assessed'*.

To summarize, the analysis of the comments provides input for recommendations for the development of the ADAPT-instrument in phase 3. First, scoring guidelines about when to give a score *cannot be assessed* must be made. Second, items and performance level descriptors need to be analysed and adjusted for ambiguity. Finally, when in phase 3 a version 3.0 of the ADAPT-instrument is developed, future raters need to be trained well to score objectively based on the performance level descriptors and how to link all data available to give a fair score.

Recommendations in Response to Phase 2. Results of phase 2 showed that all 25 items need revision except for items PV6 and LU6, and that there was a lot of difference in scores and raters often disagreed about whether a teacher's DI quality is (in)sufficient, and raters often indicated that items *could not be assessed*. Taking all results of phase 2 into account, there are a few recommendations to conduct in phase 3. One crucial step that must be made is to analyse all items and its related performance level descriptors, except for PV 6 and L1LU6, and adjust them for ambiguity. However, given the fact that only two items score a relatively high agreement, and not even full agreement, it might be wise to also analyse these items. Furthermore, a scoring rule must be developed for when to give the score *cannot be assessed* and must be added to the rater manual. When these recommendations are adapted in version 3.0 of the ADAPT-instrument, certain steps are also crucial before assessing teachers again and can be described in the rater manual. First, all information must be available in the form of a well-structured interview, based on the items and performance level descriptors of version 3.0 of the ADAPT-instrument. Second, all teacher documents have to be available, and at last the video-taped lesson(s) should be of good quality.

Furthermore, raters must first be properly trained in all the scoring rules but also in how to give a fair score, based on all the data available. Knowing that scoring the DI quality of teachers is hard because all data is different, it might be a good decision to make the comments where raters explain their given score mandatory. This gives insight in the reasoning of, and between, raters in how they gave scores. Additionally, during the training the focus per item could also be about when to score (in)sufficient per item. That is important because it would matter less if a teacher scores a 3 or 4 for one item, than when a teacher scores a 2 or 3 for one item. Scoring a 3 or 4 matter less, because both is sufficient. When raters disagree often about whether it is sufficient or not, it would raise more questions about reliability of the scores.

5.3. Results Phase 3

Together with two experts, the MATCH-trainer and the postdoctoral researcher, the researcher developed a version 3.0 of the ADAPT-instrument. During this expert meeting, the recommendations of phase 2 were discussed, the revision of items was conducted, together with

the formulation of new scoring rules. Besides, during this meeting some new recommendations for future development of the instrument were established. In this section, the new scoring rules will be described. Next, the adjustments made for the items with related performance level descriptors will be described, followed by the recommendations discussed during the meetings. At the end of this section, version 3.0 of the ADAPT-instrument will be analysed, based on the evaluation framework of Dobbelaer (2019), to understand what the next step should be in the development of this instrument.

Scorings rules. The first two pages of the ADAPT-instrument serve as a rater manual and some adjustment are made in this manual. First, the importance of the availability of the data of all phases is underlined. This is needed to give a faire score. Second, three rules were added about when to give a certain score. *(1)* A teacher should master the full content of a performance level descriptor to earn that score; *(2)* when a rater doubts between two scores, the lowest score must be chosen; *(3)* when there is information to give a teacher score 2, but no information about score 3 or 4, the rater should not fill in *could not be assessed* but give that teacher a 2 score when he or she at least meets that score.

Adjustments ADAPT-instrument. In line with what was recommended in phase 2, all items were revised because the focus group of experts thought it was better to analyse PV6 and LU6 as well. And not without result, because both items were revised in some way. Figure 9 shows how each item and performance level descriptor looks like in version 3.0 of the ADAPT-instrument.

At the start of the phase *preparation of the lesson period* it is described how a period should look like in terms of time and content. The meaning of the abbreviations, for example ‘PV1’ and ‘M’ in this case, stayed the same for all items. Furthermore, the description of the item itself (the bold section on the left) is the same as the content of score 4. In this way it is clear what each teacher should master to receive a ‘good’ per item.

Next, within the focus group the *explanatory notes* were debated because the raters found it complicated to use two large documents side by side. Therefore, the *explanatory notes* were added to the ADAPT-instrument itself and can be found underneath each performance level descriptor. When a cell of *explanatory* notes is empty, it means no further explanation is needed. The next step was to analyse and adjust each item, with the performance level descriptor and *explanatory notes*. Every item was read, discussed and adjusted to make the content of each item and/or performance level descriptor more evident and to make clear what the difference between each performance level descriptor is. The number of items in total and of items per phase remained the same. However, the focus of some items changed and elements from some

items have been ‘redistributed’. For example, items about students with a strong, weak or basis mathematics level have been clearly disassembled and placed separately. Eventually, version 3.0 of the ADAPT-instrument was developed (see Appendix I). Further in phase 3, this version 3.0 will be analysed to determine, according to evaluation framework of Dobbelaer (2019), where this instrument stands in terms of development. Based on that analysis, recommendations for further development will be given to make version 3.0 of the ADAPT-instrument more valid and reliable.

During the meetings of the focus group, recommendations for future development and /or research were discussed. The raters observed that almost every mathematics lesson starts with an automation exercise (e.g. multiplying) followed by the ‘real’ lesson and the experts recognized this component as well. Because learning to automate appears to be a crucial phase in each mathematics lesson, it should also have a place in the ADAPT-instrument. In the future, it could be investigated if and how this component can become part of the instrument.

Periodevoorbereiding

Bij een periode wordt gedacht aan een ongedefinieerd aantal weken waarbij de leerkracht de doelen in samenhang analyseert en bepaalt wat op welke momenten getoetst gaat worden.

		1	2	3	4	Score	Niet te beoordelen
PV1 M	**De leerkracht bekijkt de behaalde scores. En relateert de scores aan de inhoudelijke doelen. De leerkracht bepaalt op basis hiervan of de doelen ook inhoudelijk wel, niet of deels zijn behaald. De leerkracht gaat na waarom de doelen uit de vorige periode wel, niet, of deels zijn behaald.**	De leerkracht evalueert niet.	De leerkracht bekijkt alleen behaalde scores (bijv. niveauscore op LOVS-scores en/of cijfers op methode toetsen).	De leerkracht bekijkt de behaalde scores. En relateert de scores aan de inhoudelijke doelen. De leerkracht bepaalt op basis hiervan of de doelen ook inhoudelijk wel, niet of deels zijn behaald.	De leerkracht bekijkt de behaalde scores. En relateert de scores aan de inhoudelijke doelen. De leerkracht bepaalt op basis hiervan of de doelen ook inhoudelijk wel, niet of deels zijn behaald. De leerkracht gaat na waarom de doelen uit de vorige periode wel, niet, of deels zijn behaald.		
			Hierbij worden alleen scores en/of niveaus geconstateerd ter kennisgeving. De leerkracht kijkt in grote lijnen naar of dit naar verwachting is of niet.	*Hier wordt ook inhoudelijk naar de doelen gekeken. Een categorieanalyse, wat zowel op LOVS & methodetoetsen kan.*	*Hierbij wordt ook een verklaring gegeven voor het wel of niet behalen van de norm/inhoudelijke doelen. Een leerkracht constateert dat de plusleerlingen een bepaald doel niet hebben behaald, omdat hij/zij ze tijdens deze lessen niet aan de instructie heeft laten meedoen.*		
	Opmerkingen:						

Figure 9. Design of version 3.0 of the ADAPT-instrument, in this case from the phase ‘preparation of the lesson period’.

Evaluation Framework & ADAPT-instrument. The ADAPT-instrument has been improved, based on the results of phase 2 and in this section this version 3.0 will be analysed according to the evaluation framework of Dobbelaer (2019). As mentioned before, the

evaluation framework of Dobbelaer functions as a guide for evaluating the quality of a COS. In this phase, the evaluation framework was used to evaluate version 3.0 of the ADAPT-instrument to decide which step should be taken next in the process of further development. Table 7 shows an overview of, on the left side, the evaluation framework, and, on the right side, whether the evidence for the issues of the evaluation framework has been provided or not. Results show that version 3.0 of the ADAPT-instrument complies with the standards of the first part of the evaluation framework. The theoretical basis of the ADAPT-instrument is sufficient, and the quality of the ADAPT-instrument is expected to be good. However, a sufficient rater manual is lacking, and from there onwards evidence is missing to comply with the standards of the evaluation framework.

To summarize, phase 3 started with adjustments of version 2.0 of the ADAPT-instrument. Scoring guidelines were adjusted, the *explanatory notes* were added to the instrument, and each item was revised for clarity. This version 3.0 of the ADAPT-instrument was analysed according to the evaluation framework of Dobbelaer (2019) to understand what the next step of development should be. In the conclusion, recommendations will be given for future research, based on the results of phase 3.

Table 7
The latest version of the ADAPT-instrument in light of the evaluation framework.

Evaluation framework	
Part A. Evaluation of the relevant COTAN criteria	
1. Theoretical basis of the COS	**The ADAPT-instrument**
1.1 Is the purpose of the COS specified?	
a. Are the constructs that the COS intends to measure specified?	Yes, based on the skill hierarchy.
b. Is (are) the group(s) for which the COS is (are) intended specified?	Yes, primary school teachers.
c. Is the purpose of the COS specified?	Yes, assessing the DI quality of teachers for a mathematics lesson.
1.2 Is the theory underlying the COS described?	Yes, in this research and in the research of Keuning et al. (2017) and Van Geel et al. (2018).
1.3 Is the relevance of the COS's content for measuring the construct(s) justified?	Yes, in this research and in the research of Van Geel et al. (2018) and Keuning et al. (2017).
2. Quality of the COS materials	
2.1 Is the COS complete and clear?	That is the expectation for this improved version.

2.2 Are the items in the COS formulated correctly?	That is the expectation for this improved version.
2.3 Is the scoring system devised in such a way that scoring errors can be avoided?	That is the expectation for this improved version.
3. Quality of the rater manual	
3.1 Is a rater manual available?	Can be found on the first two pages of the instrument, but separate manual is missing.
3.2 Are the instructions for raters clear and complete?*	That is the expectation for this improved version.
3.3 Is information provided on the applications and limitations of the COS?	Decribed in this research, but not in the manual.
3.4 Is a summary of the research findings published in the manual?	Described in this research, but not in the manual.
3.5 Is the degree of expertise required by raters to use the COS specified?	Not described as a rule.
4. Norms	
4.1 Are norms provided?	No
4.2 Are the norms up to date?	N/a
Norm-referenced interpretation	
4.3 Are the norm groups large enough and representative?	N/a
Domain-referenced interpretation	
4.4 Is there sufficient agreement between raters?	Not tested for improved version
4.5 Have the raters been selected and trained appropriately?	Not yet
5. Reliability	
5.1 Is information on the reliability of the COS provided?	Not for version 3.0.
5.2 Are the findings of the reliability research sufficient considering the type of decisions that are intended to be made using the COS score?	Not for version 3.0.
5.3 What is the quality of the reliability research?	
a. Are the procedures for computing the reliability coefficients correct?	In this research, a procedure is available.
b. Are the samples for computing the reliability coefficients consistent with the intended use of the COS?	It was for this research, but not yet tested for version 3.0.
c. Is the information provided sufficient to make a substantiated judgment of the reliability of the COS?	Not yet
Part B. Evaluation of the validity argument	
The scoring inference	

Warrant 1: The scoring rule(s) is/are (statistically) substantiated	
• The COS is based on theory, research, and/or standards	Yes
• The scoring rule(s) is/are supported by experts	Yes
• The scoring rule(s) is/are supported by teachers	The hierarchy is supported by teachers, but the ADAPT-instrument not yet.
• The scoring rule(s) has/have been tested in (pilot) research	The new scoring rules not yet, the others are.
• Statistical analyses support the scoring rule(s)	No
• The psychometric quality of the items is sufficient	No
Warrant 2: Measures were taken to score accurately and consistently	
• For each criterion, scoring rules are available for raters	Yes
• Raters are trained in the use of the COS	Not for the newest version
• Raters are expected to meet a certain level of expertise	No rule yet described
• Inter-rater reliability is sufficient	N/a
• Observation scores are consistent over time	N/a
Warrant 3: Attention is dedicated to preventing rater bias	
• Attention is paid to preventing rater bias during rater training and/or in the COS manual	It is a component, before using the ADAPT-instrument, and described in this research.
• Multiple raters are used to compute an observation score	No
• Statistical analyses show that raters do not rate specific groups of teachers differently from others	No
• Inter-rater reliability is sufficient	No
The generalization inference	
Warrant: Opportunities for generalizations are explicitly described in the COS	
• The required number of teacher observations is specified and substantiated	No
• The observation length is specified and substantiated	No
• The number of raters is specified and substantiated	No
• The observation moment is specified and substantiated	No
• The lesson type is specified and substantiated	Yes, a mathematics lesson on a primary school

• A generalizability study, a reliability study, or IRT analyses has shown that the sample of observations are representative of the assessment domain	No
• Research into the variation of observed lessons supports the generalizations of the observed score	No
• Confidence intervals are available and are based on a convention	No
The extrapolation inference	
Warrant: The score in the assessment domain is related to the broader target domain	
• The assessment domain covers a great portion of the target domain (face validity)	Yes, based on the expert meetings in this study.
• The theoretical framework underlying the COS fits within the target domain	Yes
• The observed score is related to other measures within the target domain	No
The implication inference	
Warrant 1: The implications for the observable attribute are appropriate (given the theoretical construct)	Yes, based on face validity only.
Warrant 2: The implications for the observable attribute are appropriate (given the statistical analyses)	No

Note. Abbreviation N/a stands for '*not applicable*'.

6. Conclusion and Recommendations.

The aim of this study was to further develop the ADAPT-instrument, for assessing the quality of teacher's DI in primary education mathematics lessons. In the first phase, the preliminary version 1.0 of the ADAPT-instrument was developed into a version 2.0, adjusted with scoring rules in the form of a rubric with a performance level descriptor per item. In the second phase, this version 2.0 was tested for inter-rater reliability and content/face validity and based on those results an improved version 3.0 was developed in the third phase. This version 3.0 was analysed based on the evaluation framework of Dobbelaer (2019) to get insight into the next step of development.

For version 2.0, inter-rater reliability was not achieved for almost all items and, consequently, it was decided to analyse and adjust all items and related performance level descriptors within an expert meeting. Based on the findings of phase 2 and input during the expert meeting, a version 3.0 of the ADAPT-instrument was developed. This version 3.0 was

analysed with the evaluation framework of Dobbelaer (2019). Based on this analysis, it appeared that the ADAPT-instrument has sufficient theoretical basis. The quality (part one, indicator 2 of the framework) of version 3.0 of the ADAPT-instrument is expected to be good, but not tested yet. As mentioned in the framework for indicator 4.4 *Is there sufficient agreement between raters?* sufficient agreement between raters is not proven yet for version 3.0. It is, therefore, recommended to test the inter-rater reliability again, when raters are selected and trained for using version 3.0. The results of this inter-rater reliability will then again provide feedback for the previous part about the quality of the ADAPT-instrument. It might turn out, for example based on the comments of raters, that the items are not formulated clear, and the scoring system needs revision. However, the expectation is positive in this way that, after the adjustments made in phase 3, the formulation of the items and related performance level descriptors is less ambiguous and scoring errors are avoid more. Additionally, for future usage of the ADAPT-instrument, a few points to focus on during rater training occurred to be important. Raters must be trained in all the scoring rules, in the difference between sufficient and insufficient per item, and how to use all data available to come to an overall score. In future research this should turn out positively for the inter-rater agreement about whether a teacher's quality of DI is (in)sufficient.

In addition, results of phase 2 showed that the explanation for differences in scores could be categorized into three causes: missing or incomplete scoring rules, misunderstanding of scoring rules, and an incomplete manual. These recommendations were taken in account for developing version 3.0 and included in the first two pages, which serve as a rater manual. However, Dobbelaer (2019) states it is important to establish a rater manual to gain sufficient inter-rater reliability and indicates it might be better to develop an extended rater manual for the ADAPT-instrument separate from the instrument. This result is underlined in the findings, described in Table 7 of phase 3, where the quality of the rater manual is not achieved. One of those criteria of evidence is the rule that a rater needs to meet a certain level of expertise. This is missing in the current rater manual. All raters in this study graduated from a teacher training academy (PABO) for primary education and this might be the level of expertise needed. However, one could question whether a rater also needs to have work experience as a primary school teacher. It might also be that raters must meet other criteria, for example criteria that have nothing to do with education. Overall, considering the ADAPT-instrument as a tool to assess the quality of teaching, it suggests that a rater should have some expertise in the field of education. Differences in ratings across rater with various backgrounds could be investigated in future research, in order to determine more specific rater characteristics.

In addition, when the recommendations mentioned before are achieved, this should provide new evidence for the criteria of the scoring inference in the evaluation framework of Dobbelaer (2019). Scoring inference, according to Dobbelaer, is about building an argument that a sample of an observation can be connected to an observed score. This argument can be built with gaining inter-rater reliability and well supported, tested, and trained scoring rules, as recommended.

Furthermore, findings show adjustments are also needed in the collection of data which are not mentioned in the evaluation framework of Dobbelaer (2019). In phase 2, when analysing all data gathered from raters, it appeared that information was lacking in order to give a score to an item. This was evidenced from the large number of scores *could not be assessed* and underlined in the comments of raters. Therefore, another recommendation is to adjust the interview guidelines, used to question teachers about the (preparation of the) video-taped lesson, to make sure all items are covered.

At last, another recommendation was mentioned during the expert meeting in phase 3, which entails the fact that a lot of mathematics lesson starts with an automation exercise before starting the 'real' lesson. However, there is no item yet to assess the quality of DI for this topic. Future research could investigate if and in which way this topic should be implemented in the ADAPT-instrument.

7. Discussion and Evaluation

Based on the CTA of Keuning et al. (2017) and Van Geel et al. (2019), it is concluded that the competency DI can be distinguished in four phases, which occur before, during and after the lesson. The ADAPT-instrument is developed to assess this competency DI and in order to increase truth value (Boudah, 2011) of this instrument, multiple assessment methods (e.g. lesson observation, interview, document analysis) were integrated in the instrument to come to an aggregated score. However, like Dobbelaer (2019) stated, completely objective scores are impossible to achieve. In this study the comments of raters showed different interpretations of data and inter-rater reliability was not achieved. Given these results; knowing that mathematics lessons are never the same; and the statement that DI is a very complex teaching skill (Dixon, Yssel, McConnel, & Hardin, 2014; Eysink, Hulsbeek, & Gijlers, 2017; George, 2005; Grift, Wal, & Torenbeek, 2011; Keuning et al., 2017; et al., 2018; Van Geel et al., 2018), the idea that assessing a professional competency is very complex is confirmed (Baartman et al. 2016).

When raters explained their scores in the section for comments, this gave the insight needed to understand the reasoning of raters to give a certain score. Comments of raters showed

that some items of the phase *teacher adequately address the differences between students during the lesson* could (only) be assessed due to information from the preparation phases. This emphasizes the statement of Van Geel et al. (2018) that teacher's acting during a lesson is based on choices which are made before, during, and after instruction and therefore their rationales should be taken in account. Also, it underlines that comments of raters are needed to understand where information came from, and to understand the rater in the way he or she gave the score. Besides that, it underlines the importance of data gathered from the interview with the teacher. For assessing DI, more information is needed about what cannot be seen; the reasoning of the teachers. This information is not easy to assess, because it means those items must be discussed thoroughly, and a risk is that teachers might present themselves different than they are.

Besides that, questions can be raised about the feasibility of the ADAPT-instrument. To give a score for all 27 items, a rater must carefully analyse the video-taped lesson(s), the interview, and the teacher documents (if available). All these added guidelines, research methods, and rubrics will, according to Dobbelaer (2019), reduce subjectivity as much as possible. Nevertheless, completing the ADAPT-instrument is a time-consuming job and the question is, even when it would be reliable and valid according to the evaluation framework, whether it is useable for example high-stake summative purposes. Like mentioned in the introduction, the Dutch Inspectorate of Education (2018) currently uses only four items to assess the quality of DI in primary schools. The question would be if they, instead of those four items, will be able to use the ADAPT-instrument because, like Van Geel et al. (2018) already foresaw, it would require much time and effort from skilful assessor(s).

On the contrary, information needs to be very robust to make high stake decisions (Dijkstra et al., 2012), and if that is not the case it would bring the reliability and validity of scores into question (Dobbelaer, 2019). And so, whether it's about formative or summative purposes it suggests that, to assess professional competencies like DI in a valid and reliable way, it cannot be determined without much time and effort from raters. The issue here is whether organisations, for example a Dutch Inspectorate of Education, have the discipline and/or the priority in doing this. Apart from that, according to Van Vleuten (2016) assessment should be about assessing *for* learning and not *of* learning. In the case of the Netherlands, where teachers need to develop their DI skills (Dutch Inspectorate of Education, 2018), this might be worth all time and effort needed.

In phase 3 of this study, all items of version 2.0 of the ADAPT-instrument were adjusted to make the content of items and the performance level descriptors less ambiguous. To achieve this, the content is reduced and became more concrete. This might help to reach a proper inter-

rater reliability, but this does not mean that the items are still valid (Carmines & Zeller, 1979; Kirk & Miller, 1986; Vos, 2009). In first instance, the ADAPT-instrument needs to reflect the differentiations skill hierarchy of Van Geel et al. (2018). However, the question arises whether the instrument still covers the content of this model after all the adjustments made. In the first and third phase of this study, experts in the field were involved to make sure the constructs and items were, based on theory, well defined and clear. In this light, evidence for the content and face validity is gathered (Vos, 2009) and consequently, according to the framework for construct of validity of Trochim and Donnelly (2006), translation validity is ensured. The next step might be to investigate whether criterion validity (relational approach) could be ensured. However, when in the future it appears that the ADAPT-instrument needs revision of items and performance level descriptors again, future research should therefore again focus on translation validity first.

To conclude, in this study the preliminary version 1.0 of the ADAPT-instrument has undergone major development, which resulted into an improved version 3.0 of the ADAPT-instrument. Dobbelaer (2019) states that in a developing process it is important to decide which evidence for reliability and validity is most important in their own specific situation. Information given above give directions for the next steps to take. Concluding that assessing DI is very complex, does not mean that is impossible and that the development of this instrument should stop. On the contrary, a lot of steps have been taken to make the ADAPT-instrument more valid and reliable and there are high expectations that the next criteria of the evaluation framework of Dobbelaer (2019) will be met. Furthermore, the whole process of development in this study gives a positive prospect that, in the end, there will be a valid and reliable instrument for Assessing Differentiation in All Phases of Teaching.

References

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. Studies in Educational Evaluation, 32(2), 153–170. doi:10.1016/j.stueduc.2006.04.006

Bosker, R. J. (2005). De grenzen van gedifferentieerd onderwijs [The limits of differentiated instruction] (Oration). Retrieved from https://www.rug.nl/research/portal/files/14812458/bosker.pdf

Boudah, D. J. (2011). *Conducting Educational Research: Guide to Completing a Major Project.* London: Sage.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and Validity Assessment.* http://dx.doi.org.ezproxy2.utwente.nl/10.4135/9781412985642

Dijkstra, J., Galbraith, R., Hodges, B. D., McAvoy, P. A., McCrorie, P, Southgate, L. J., … Schuwirth, L. W. T. (2012). Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education, 12*(20), 1-8. https://doi.org/10.1186/1472-6920-12-20

Dixon, A., Yssel, N., McConnel, J. M., & Hardin, T. (2014). Differentiated Instruction, Professional Development, and Teacher Efficacy. *Journal for the Education of the Gifted, 37*(2), 111-127. doi: 10.1177/0162353214529042

Dobbelaer, M.J. (2019). *The quality and qualities of classroom observation system* (Doctoral dissertation). Enschede: Ipskamp.

Eysink, T. H. S., Hulsbeek, M., & Gijlers, H. (2017). Supporting primary school teachers in differentiating in the regular classroom. *Teaching and Teacher Education, 66*, 107-116. http://dx.doi.org/10.1016/j.tate.2017.04.002

George, P. S. (2005). A Rationale for differentiating instruction in the regular classroom. *Theory Into Practice*, *44*(3), 185–193. http://doi.org/10.1207/s15430421tip4403_2

Grift, van den, W., Wal, van der, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogische didactische vaardigheid van leraren in het basisonderwijs. *Pedagogische studiën, 88,* 416-432.

Hernon, P., & Schwartz, C. (2009). Reliability and validity. *Library & Information Science Research, 31,* 73-74. doi:10.1016/j.lisr.2009.03.001

Inspectie van het Onderwijs (2018). De staat van het primair onderwijs: Onderwijsverslag 2016/2017 [The state of education in the Netherlands: the 2016/2017 Education

Report]. Utrecht, The Netherlands. Retrieved from https://www.onderwijsinspectie.nl/onderwerpen/staat-van-het-onderwijs/documenten/rapporten/2018/04/11/rapport-de-staat-van-het-onderwijs

Inspectie van het Onderwijs (2018). Naar een nieuw onderwijsresultatenmodel primair onderwijs. Retrieved from https://www.onderwijsinspectie.nl/onderwerpen/onderwijsresultaten-primair-onderwijs/naar-een-nieuw-onderwijsresultatenmodel

Inspectie van het Onderwijs (2018). Betrouwbaarheid en fairness van het inspecteursoordeel. Rapportage fairnessonderzoek 2017. Retrieved from https://www.onderwijsinspectie.nl/onderwijssectoren/primair-onderwijs/documenten/rapporten/2018/02/08/betrouwbaarheid-en-fairness-van-het-inspecteursoordeel

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement 4th edition (pp. 17– 64). Westport, Ireland: American Council on Education and Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73. https://doi.org/10.1111/jedm.12000

Keuning, T., Geel, van, M., Frèrejean, Merriënboer, van, J., Dolmans, D., & Visscher, A. J. (2017). Differentiëren bij rekenen: een cognitieve taakanalyse van het denken en handelen van basisschoolleerkrachten. *Pedagogische studiën*, *94*, 160-181.

Kirk, J., & Miller, M. L. (1986). *Reliability and Validity in Qualitative Research.* http://dx.doi.org.ezproxy2.utwente.nl/10.4135/9781412985659

Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., ... Allen, M. (2018). Teachers' instructional adaptations: A research synthesis. *Review of Educational Research*, *88*(2), 205–242. doi:10.3102/0034654317743198

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Roy, A., Guay, F., & Valois, P. (2013). Teaching to address diverse learning needs: Development and validation of a Differentiated Instruction Scale. *International Journal of Inclusive Education*, *17*(11), 1186–1204. doi:10.1080/13603116.2012.743604

Trochim, W. M., & Donnelly, J. P. (2006). The research methods knowledge base (3rd ed.). Cincinnati, OH: Atomic Dog.

Van der Vleuten, C. P. M. (2016). Revisiting "Assessing professional competence: From

methods to programmes." Medical Education, 50(9), 885–888. doi:10.1111/medu.12632

Van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & Van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. Medical Teacher, 34(3), 205–214. doi:10.3109/0142159X.2012.652239 Vos, 2009

Van Geel, M., Keuning, T., Frèrejean, J., Dolmans, D., Van Merriënboer, J., & Visscher, A. (2018). Capturing the complexity of differentiated instruction. School Effectiveness and School Improvement, 1744-5124. https://doi.org/10.1080/09243453.2018.1539013

Appendices

Appendix A. Version 1.0 ADAPT-instrument

(confidential, removed for public version)

Appendix B. Version 1.0 of the explanatory notes

(confidential, removed for public version)

Appendix C. Interview questions

(confidential, removed for public version)

Appendix D. Version 2.0 ADAPT-instrument

(confidential, removed for public version)

Appendix E. Version 2.0 explanatory notes

(confidential, removed for public version)

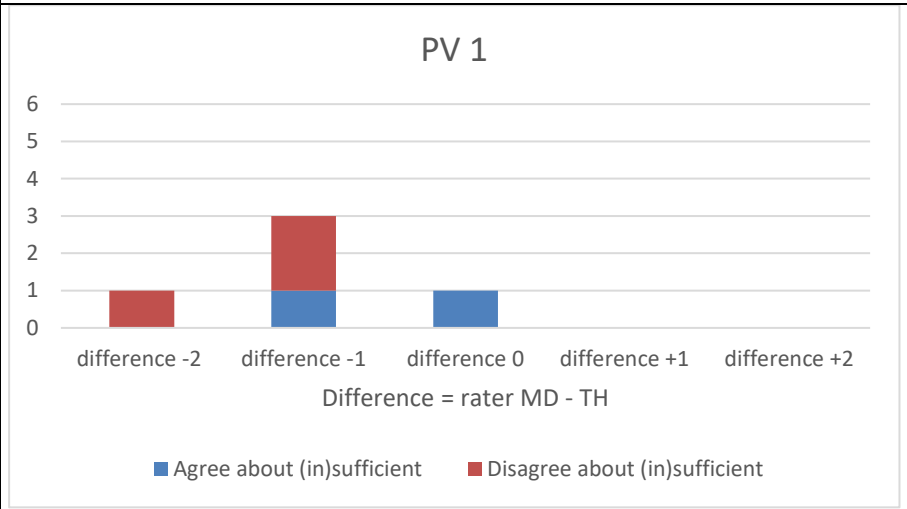
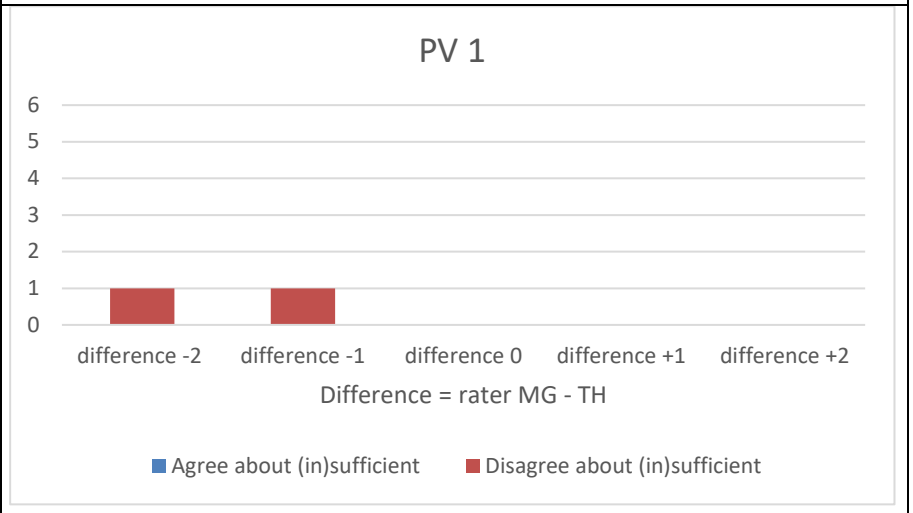
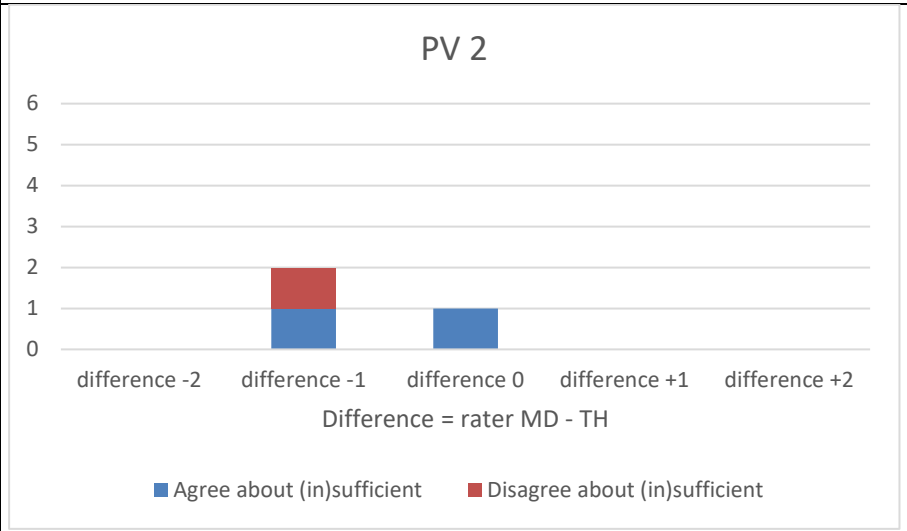
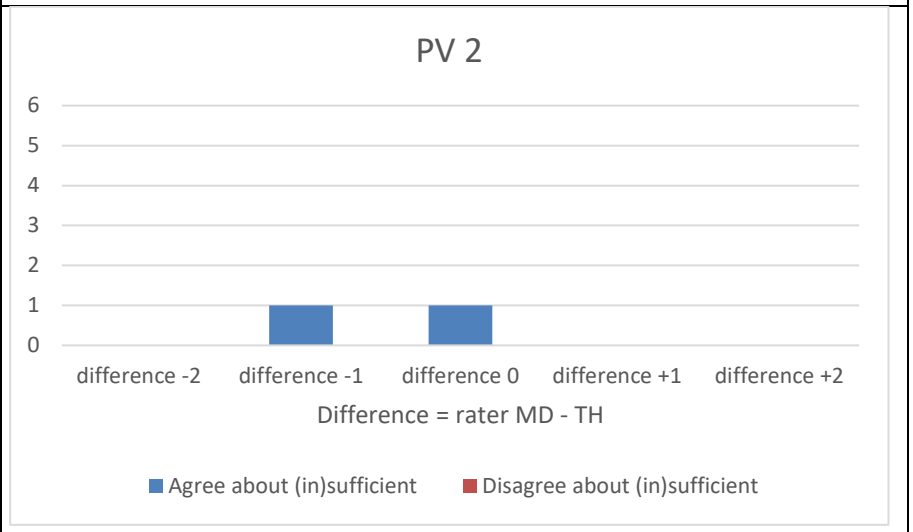
Appendix F. Scoring form

(confidential, removed for public version)

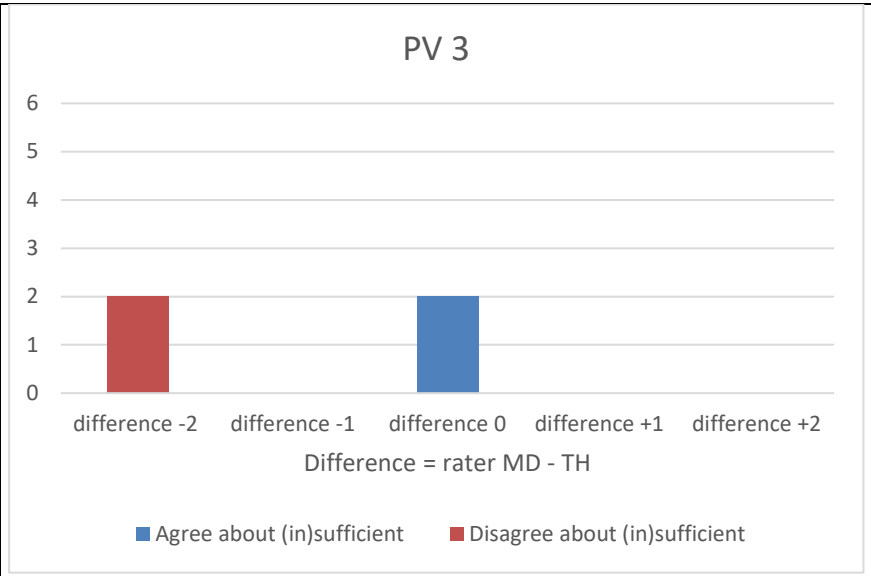
Appendix G. Raw scores of MD-TH & MG-TH

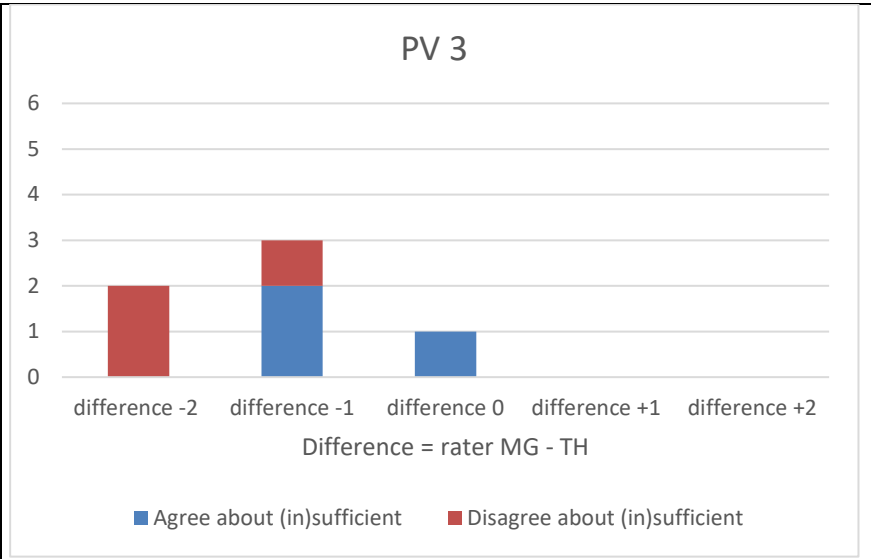
(confidential, removed for public version)

Appendix H. Difference in scores of MD-TH & MG-TH

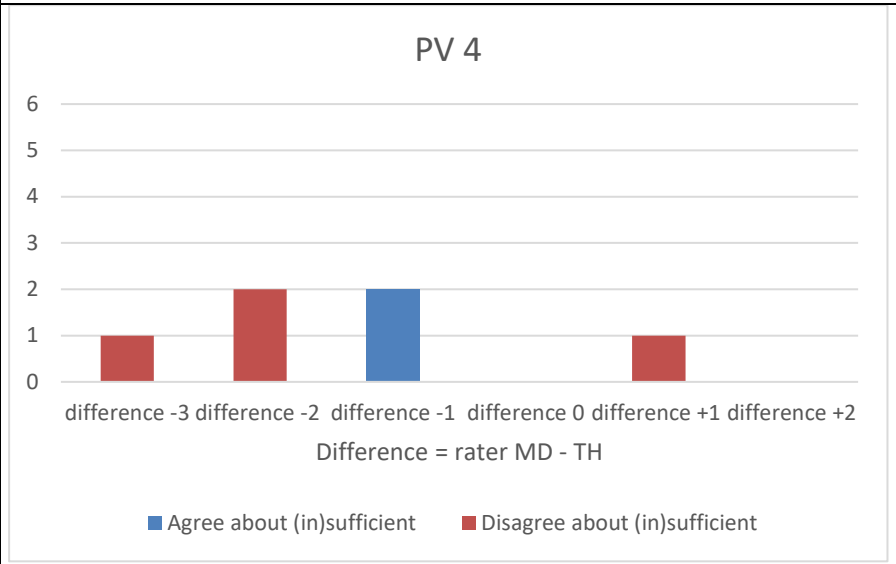
	MD – TH	MG - TH
PV 1		
PV2		

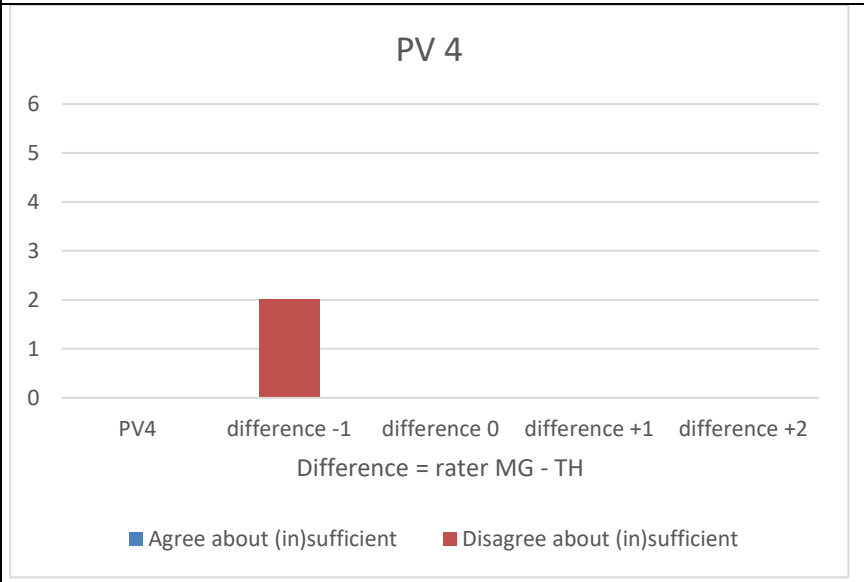
PV3


PV 3
6
5
4
3
2
1
0
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MD - TH
Agree about (in)sufficient
Disagree about (in)sufficient

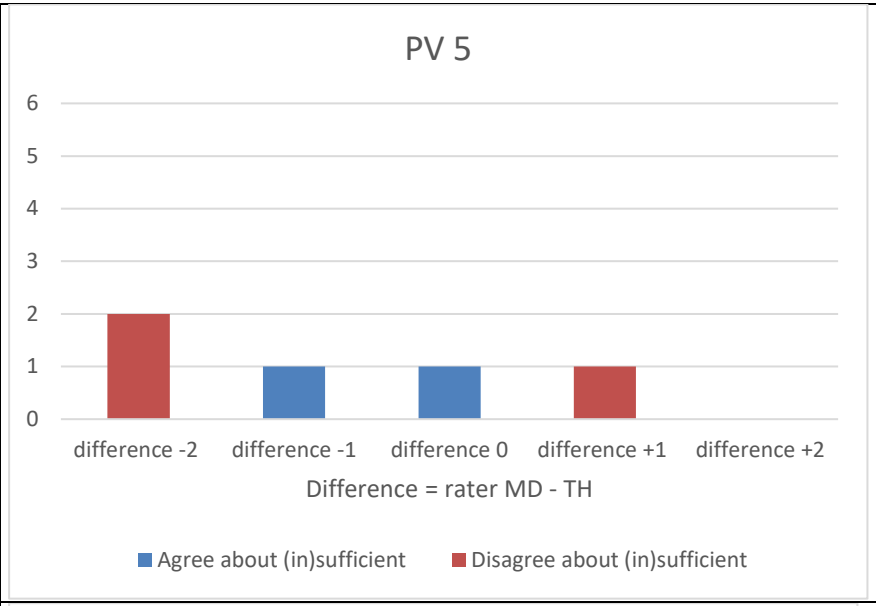

PV 3
6
5
4
3
2
1
0
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MG - TH
Agree about (in)sufficient
Disagree about (in)sufficient

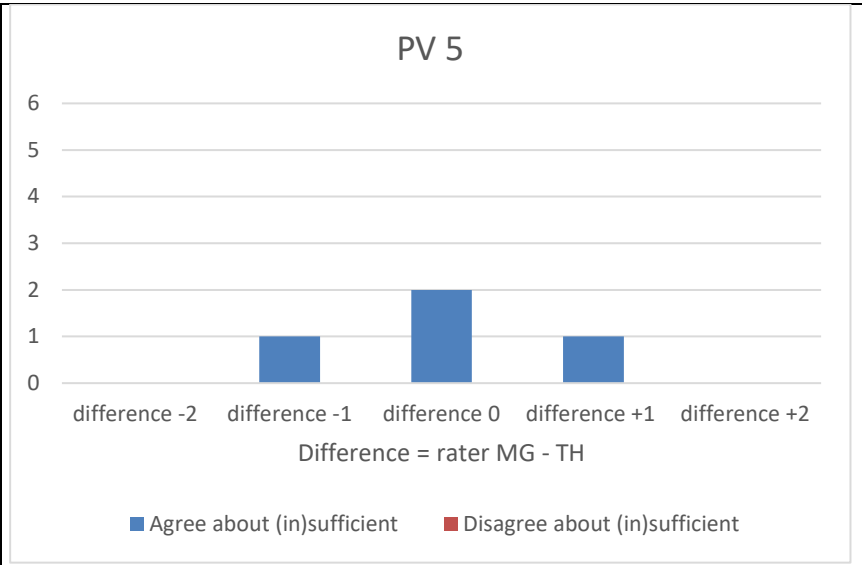
PV4


PV 4
6
5
4
3
2
1
0
difference -3
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MD - TH
Agree about (in)sufficient
Disagree about (in)sufficient

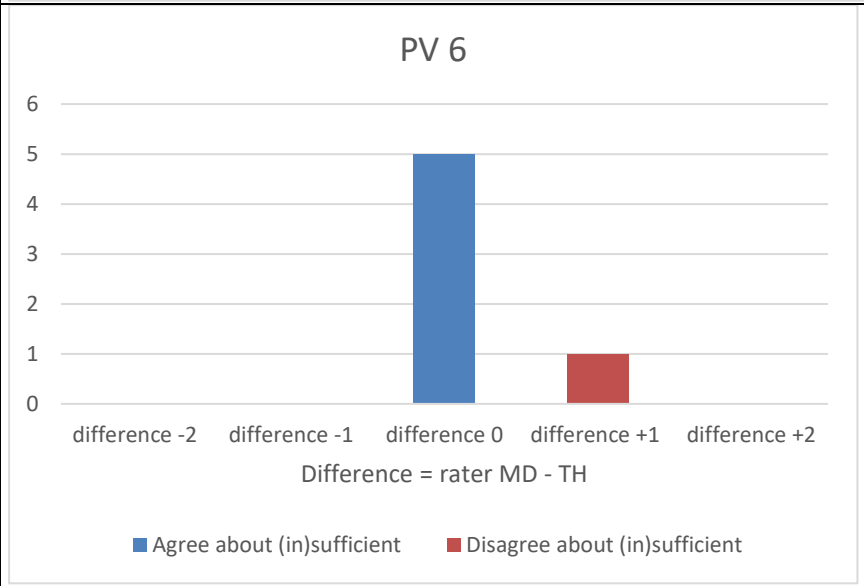

PV 4
6
5
4
3
2
1
0
PV4
difference -1
difference 0
difference +1
difference +2
Difference = rater MG - TH
Agree about (in)sufficient
Disagree about (in)sufficient

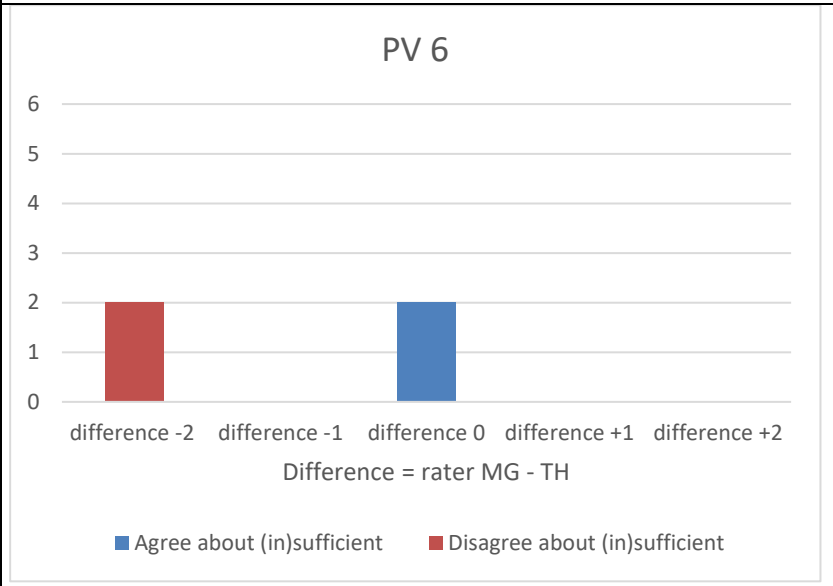
PV5


PV 5
6
5
4
3
2
1
0
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MD - TH
Agree about (in)sufficient
Disagree about (in)sufficient

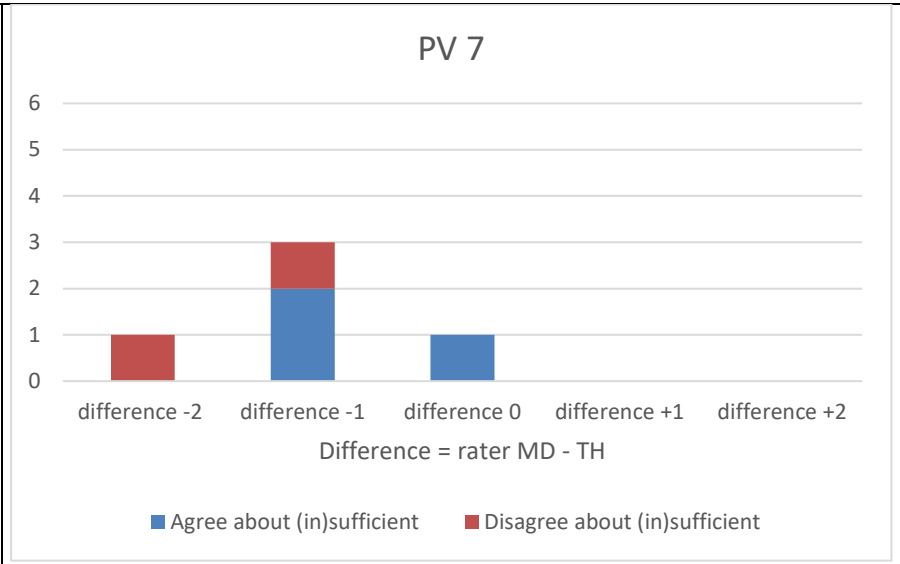

PV 5
6
5
4
3
2
1
0
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MG - TH
Agree about (in)sufficient
Disagree about (in)sufficient

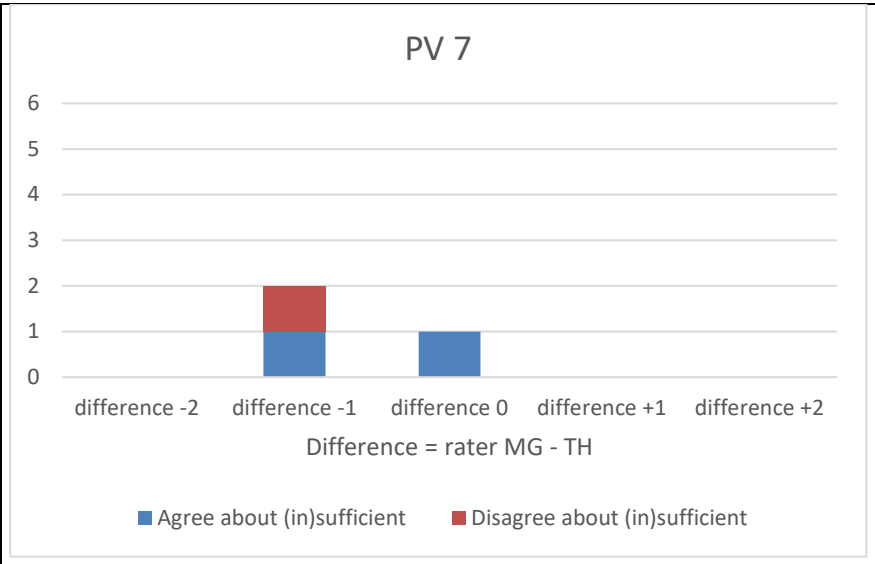
PV6


PV 6
6
5
4
3
2
1
0
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MD - TH
Agree about (in)sufficient
Disagree about (in)sufficient

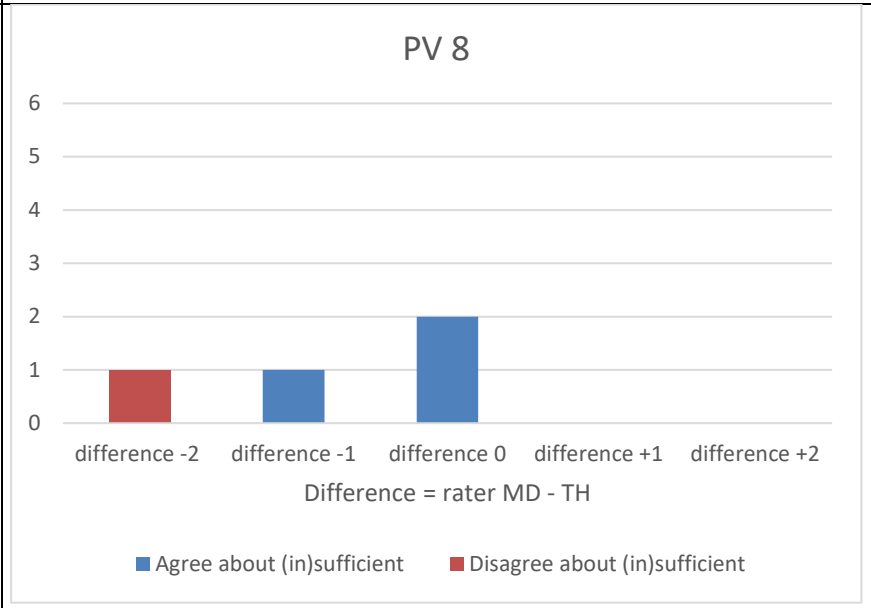

PV 6
6
5
4
3
2
1
0
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MG - TH
Agree about (in)sufficient
Disagree about (in)sufficient

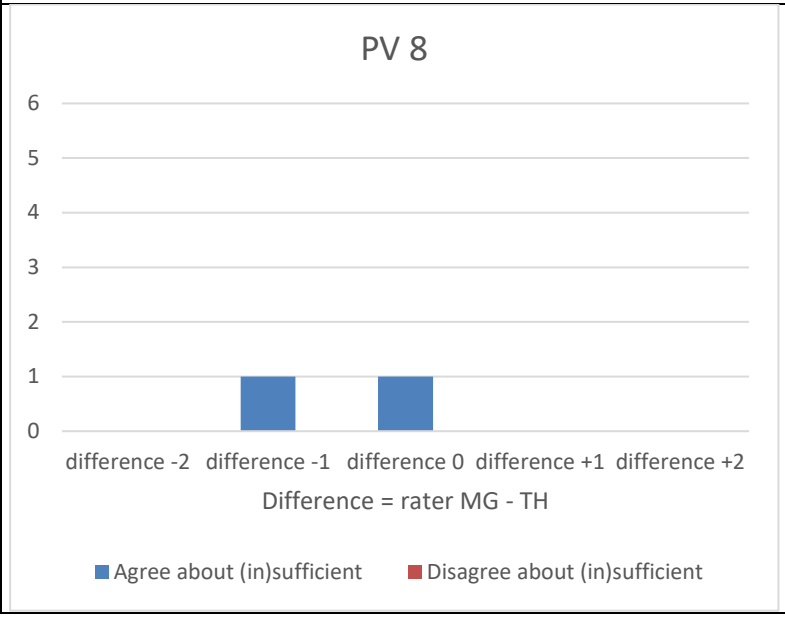
PV7

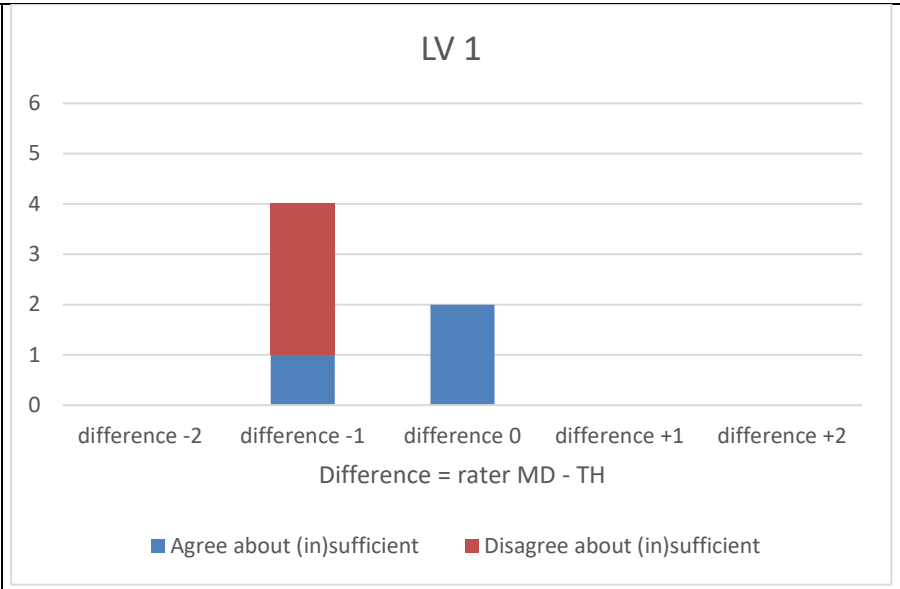
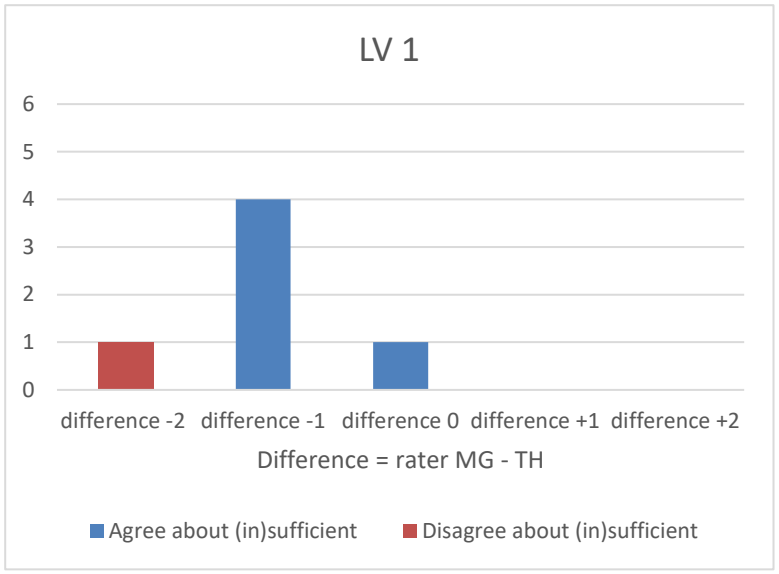
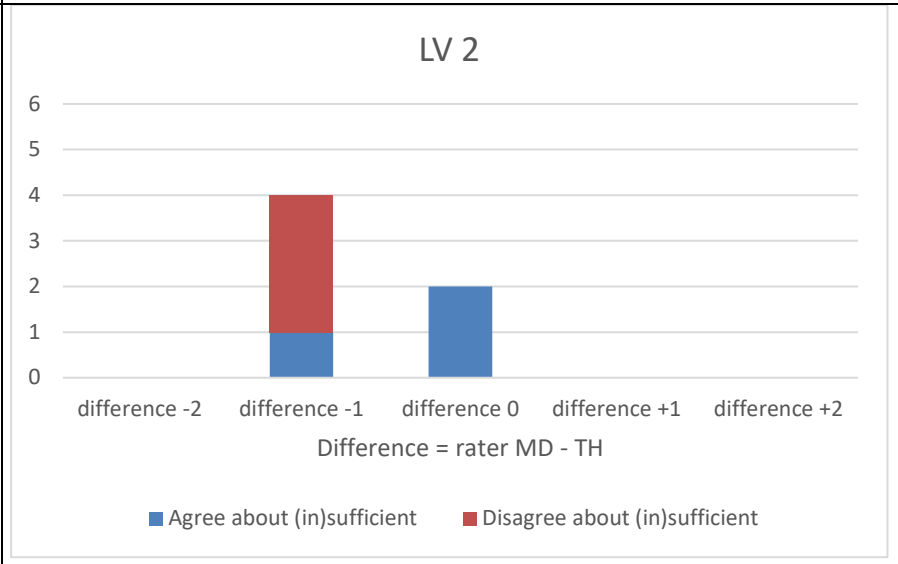
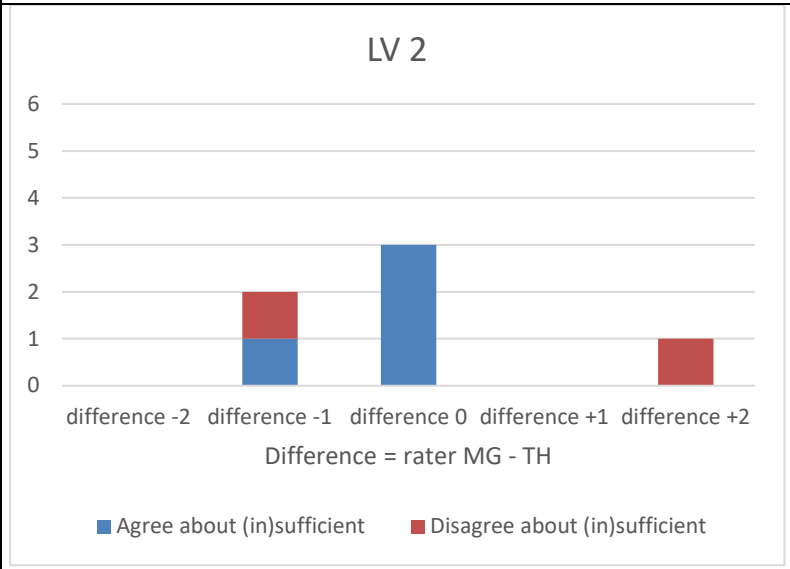

PV 7
6
5
4
3
2
1
0
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MD - TH
Agree about (in)sufficient
Disagree about (in)sufficient

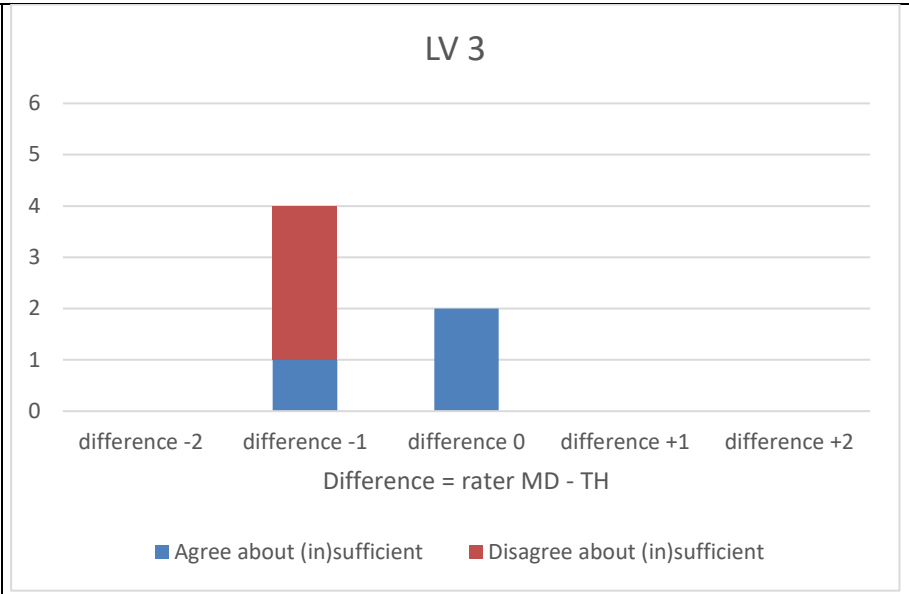
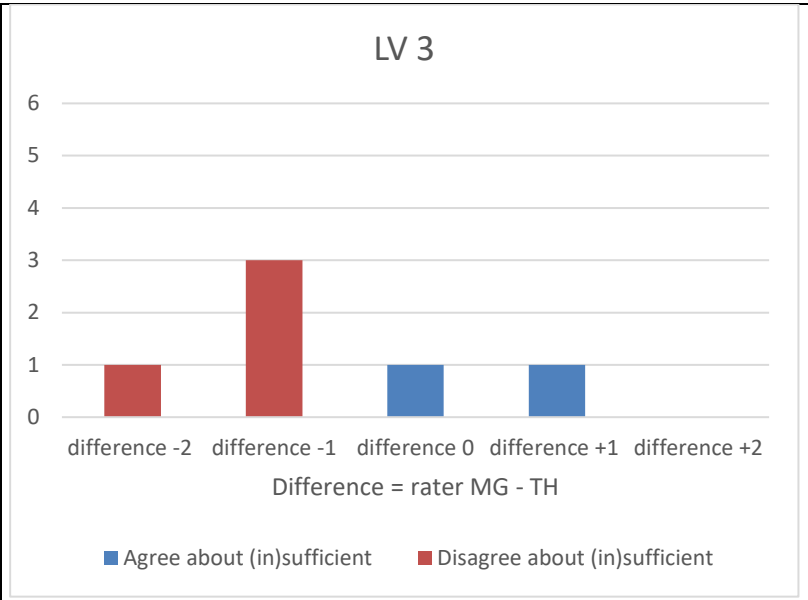
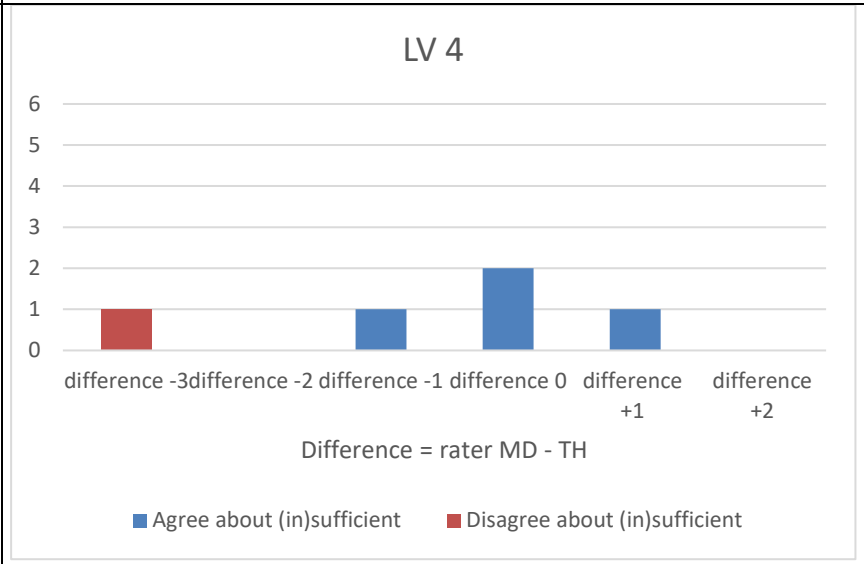
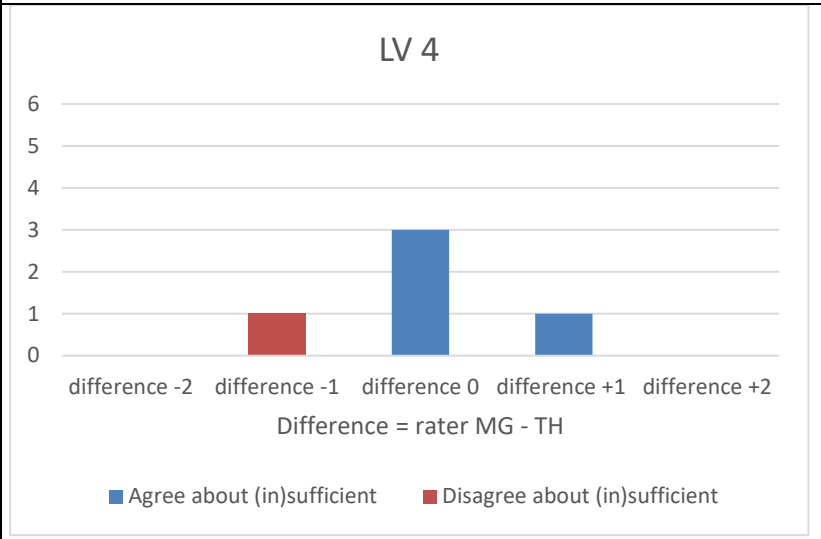

PV 7
6
5
4
3
2
1
0
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MG - TH
Agree about (in)sufficient
Disagree about (in)sufficient

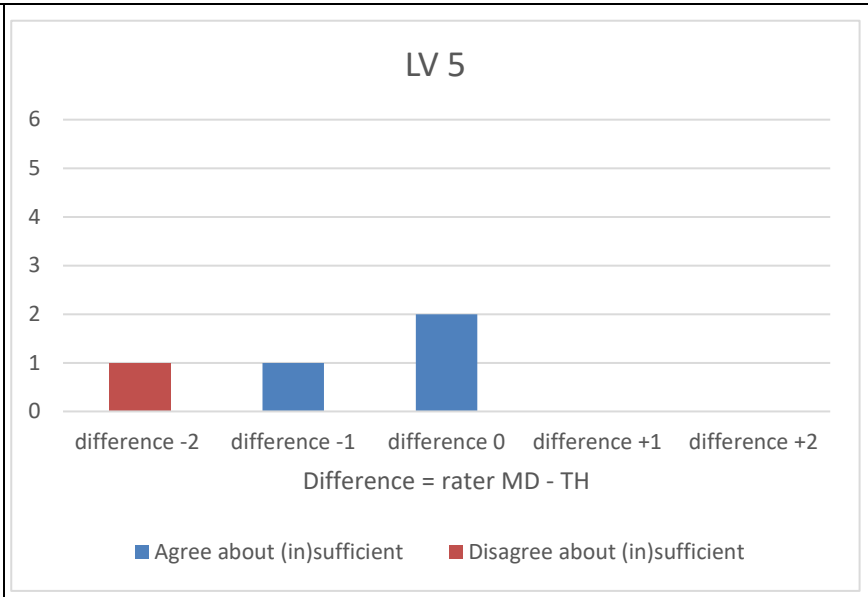
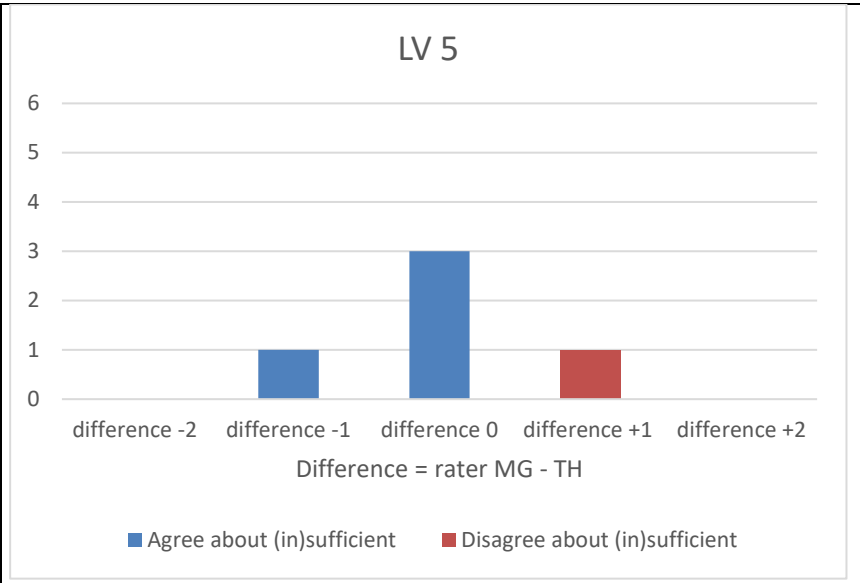
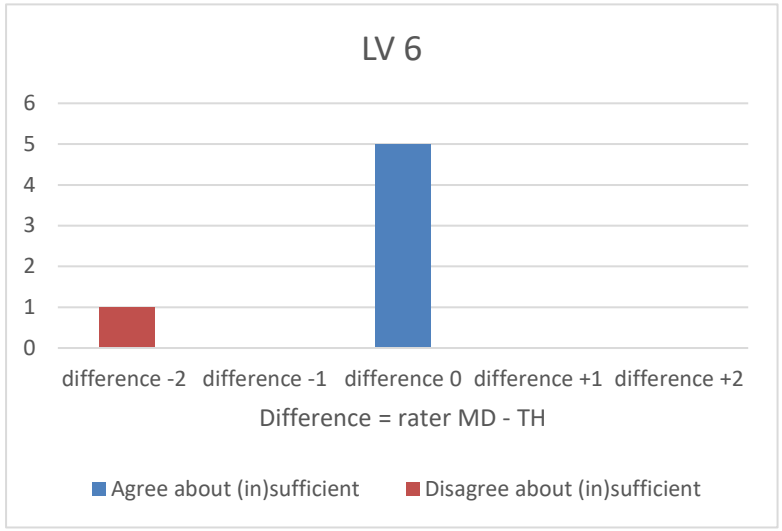
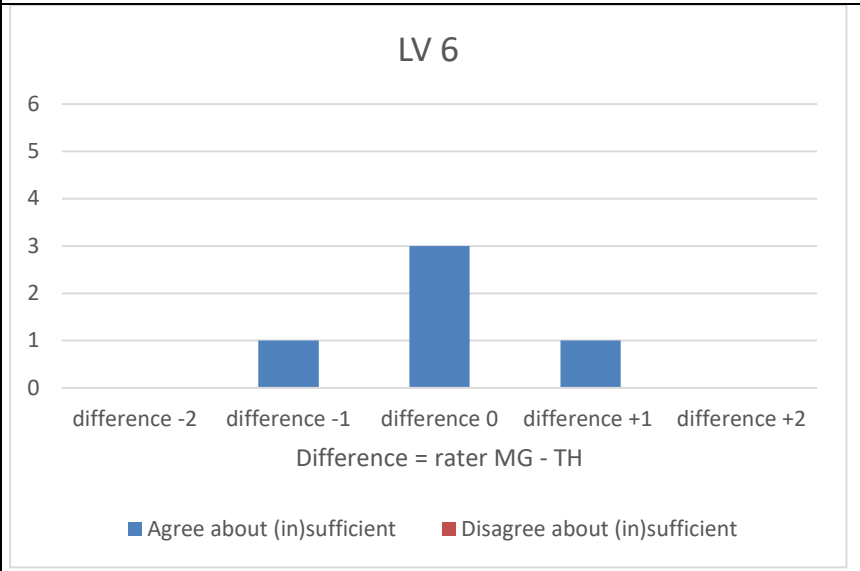
PV8

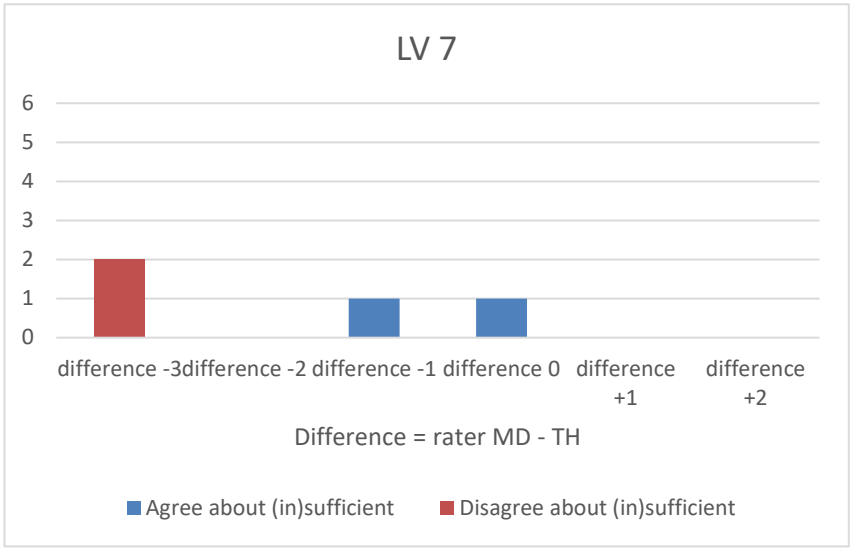
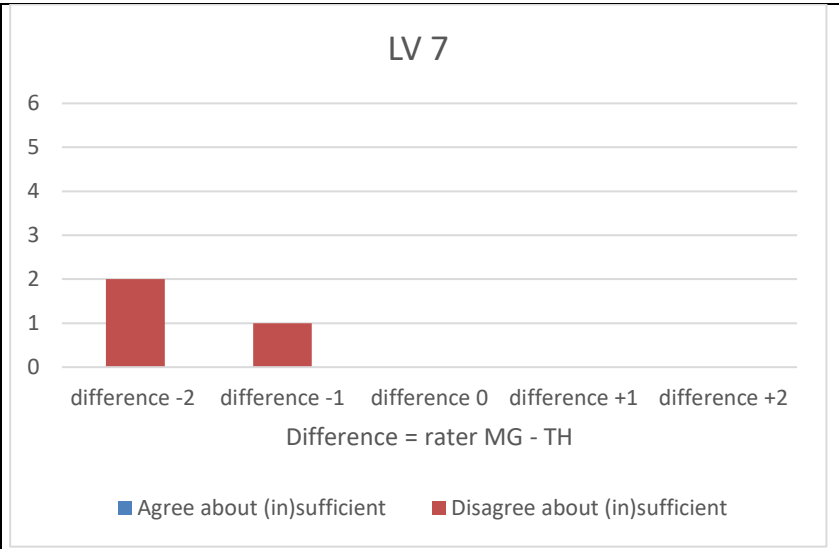
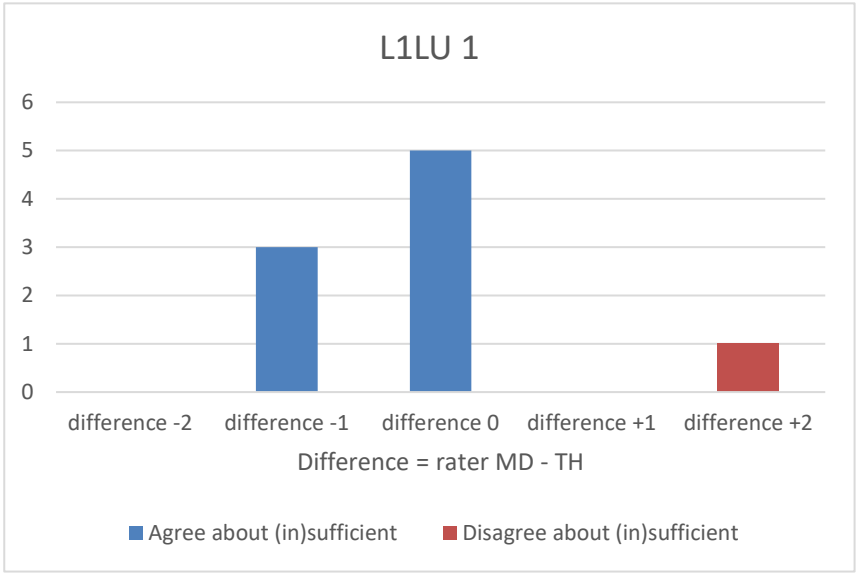
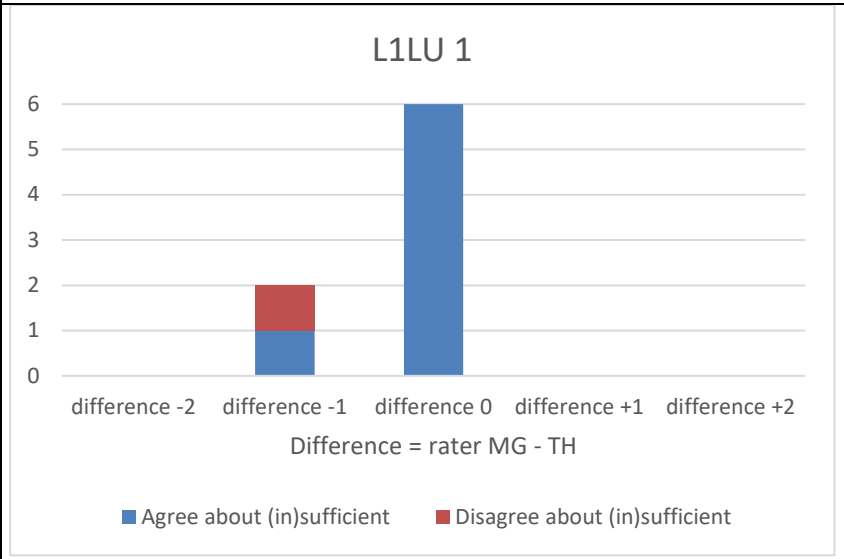

PV 8
6
5
4
3
2
1
0
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MD - TH
Agree about (in)sufficient
Disagree about (in)sufficient

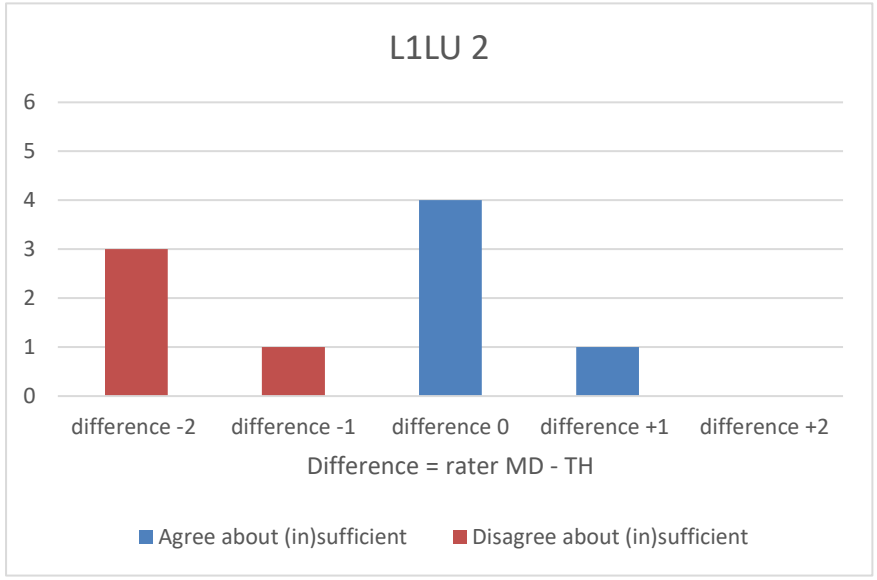
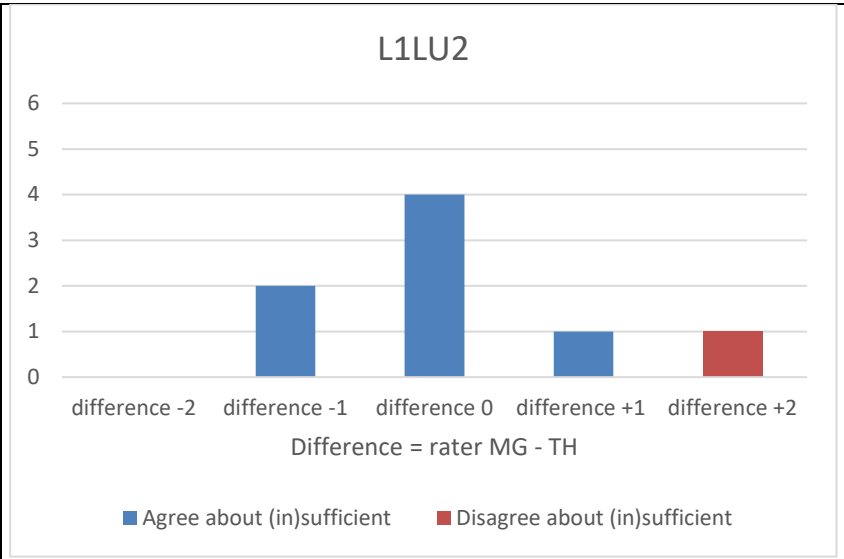
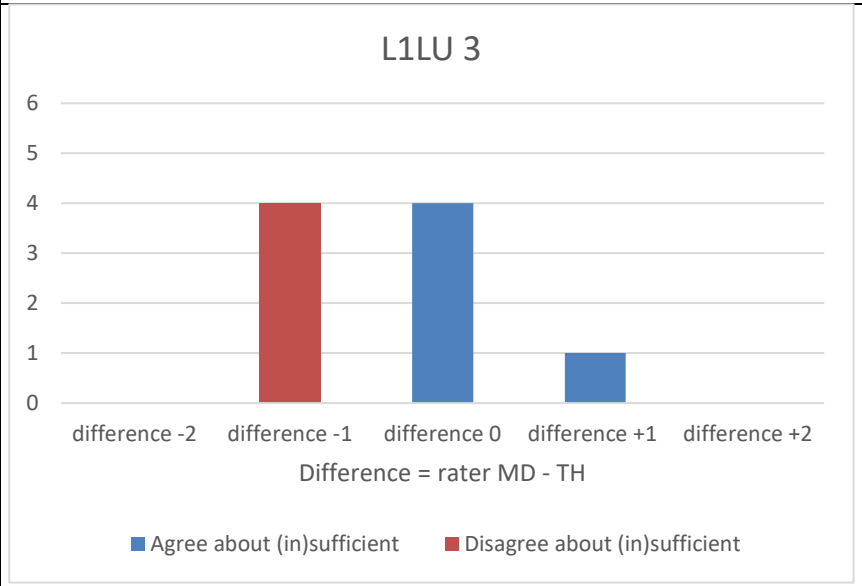
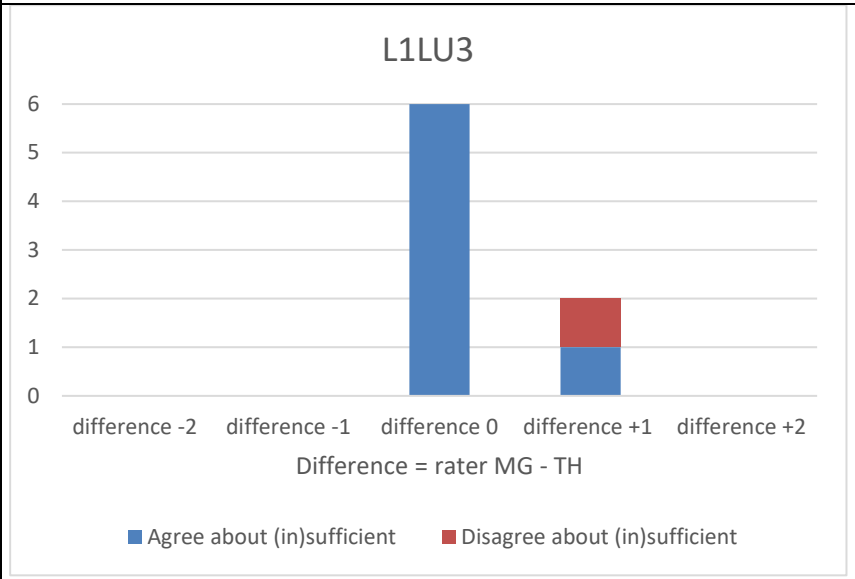

PV 8
6
5
4
3
2
1
0
difference -2
difference -1
difference 0
difference +1
difference +2
Difference = rater MG - TH
Agree about (in)sufficient
Disagree about (in)sufficient

LV1	 LV 1 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 LV 1 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient
LV2	 LV 2 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 LV 2 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient

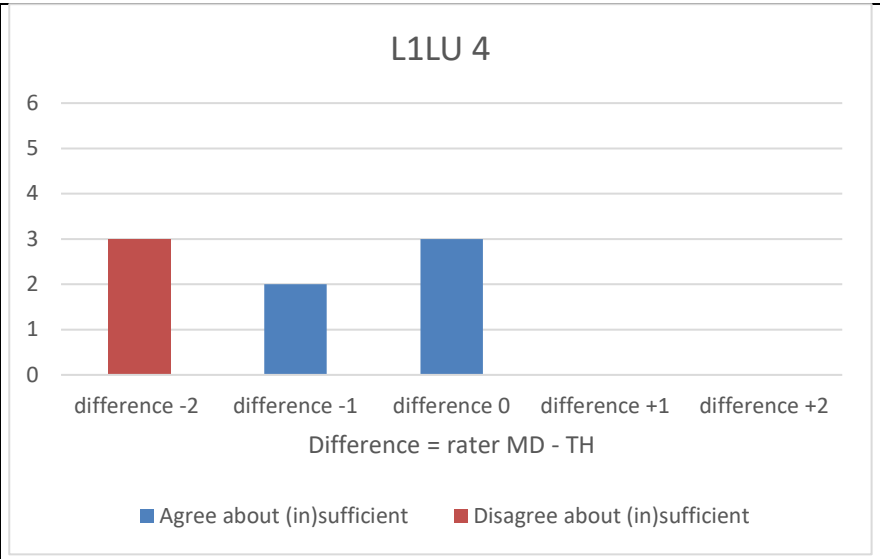
LV3	 LV 3 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 LV 3 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient
LV4	 LV 4 6 5 4 3 2 1 0 difference -3 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 LV 4 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient

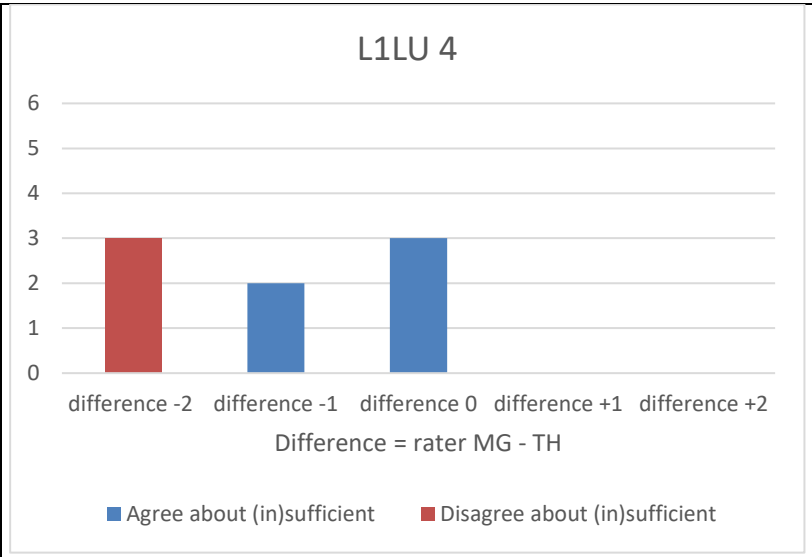
LV5	 LV 5 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 LV 5 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient
LV6	 LV 6 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 LV 6 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient

LV7	 LV 7 6 5 4 3 2 1 0 difference -3 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 LV 7 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient
L1LU1	 L1LU 1 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 L1LU 1 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient

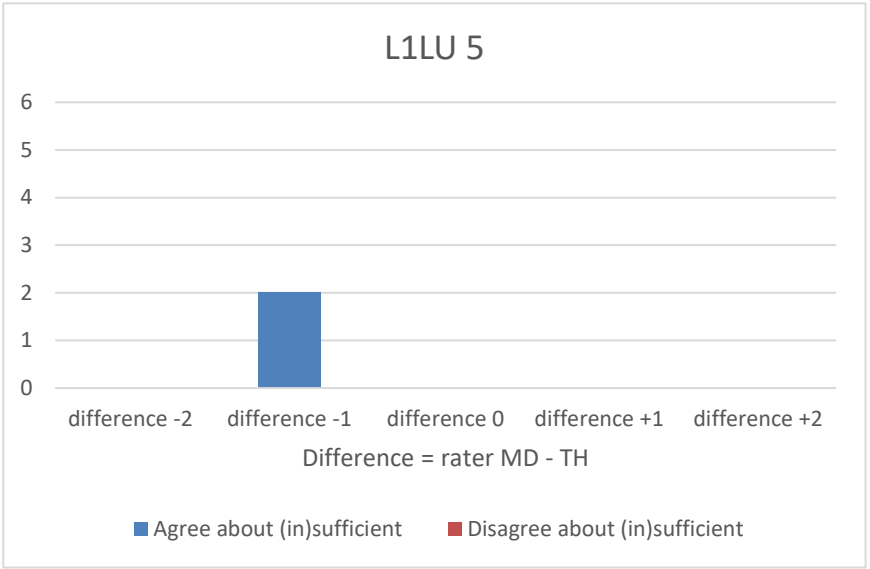
L1LU2	 L1LU 2 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 L1LU2 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient
L1LU3	 L1LU 3 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 L1LU3 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient

L1LU4

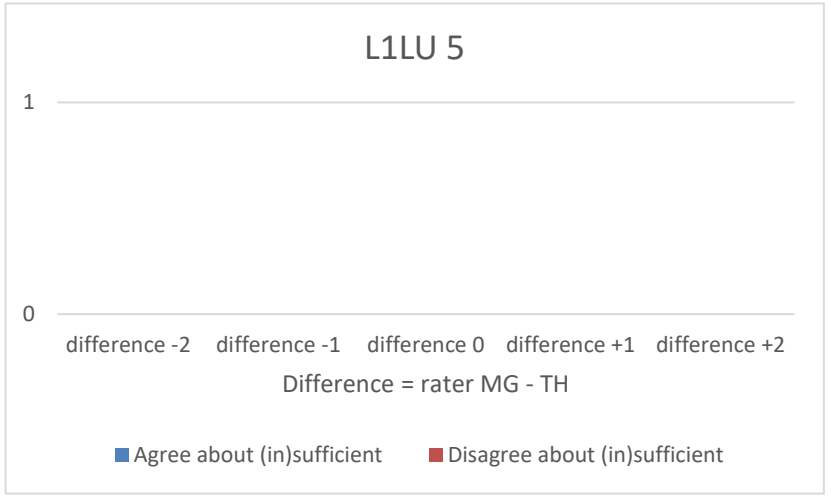


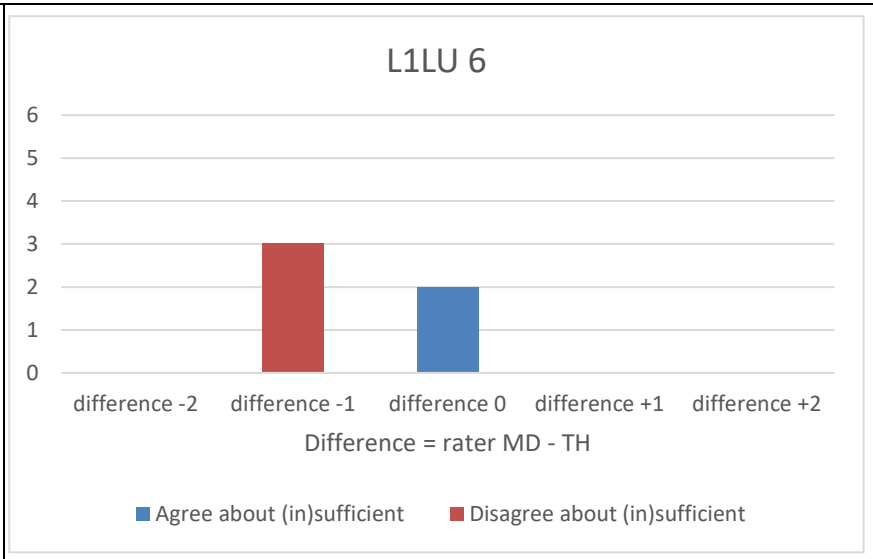
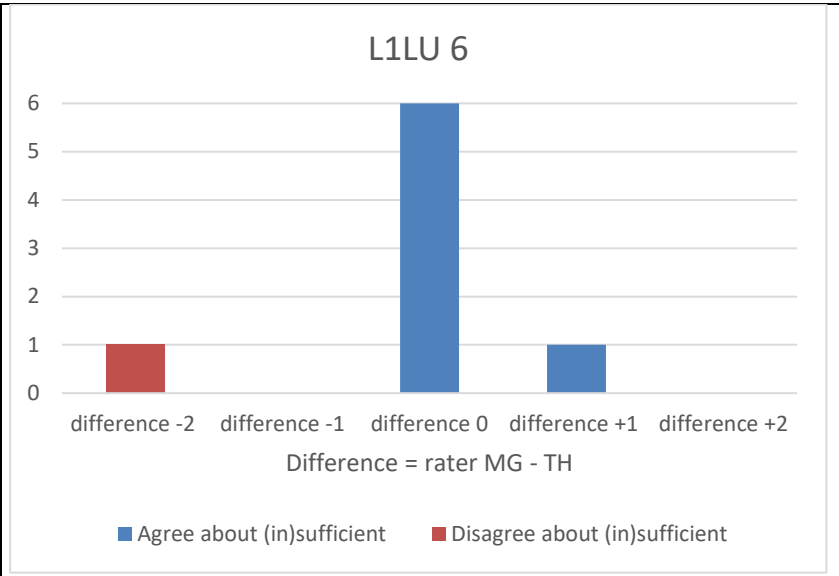
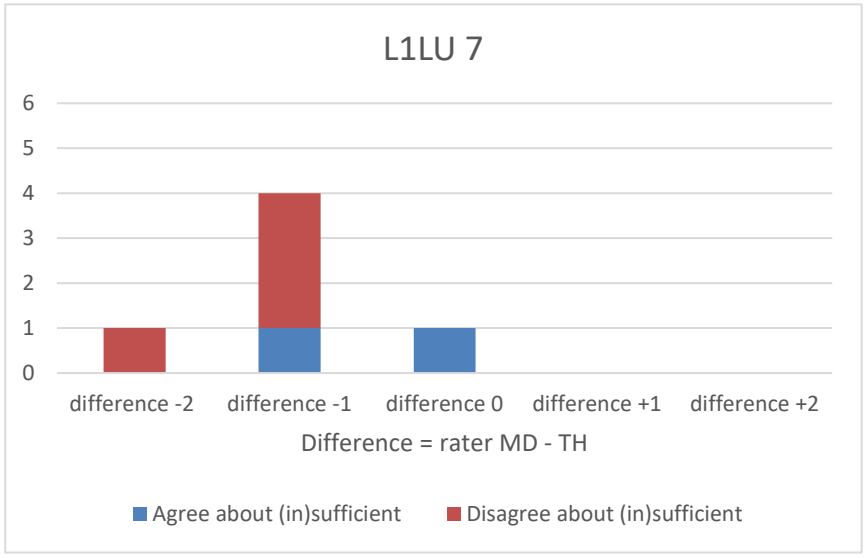
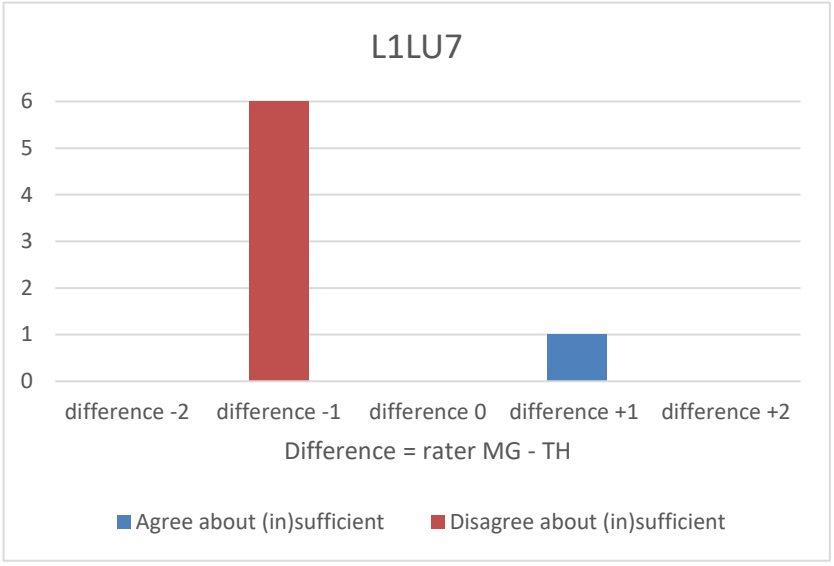


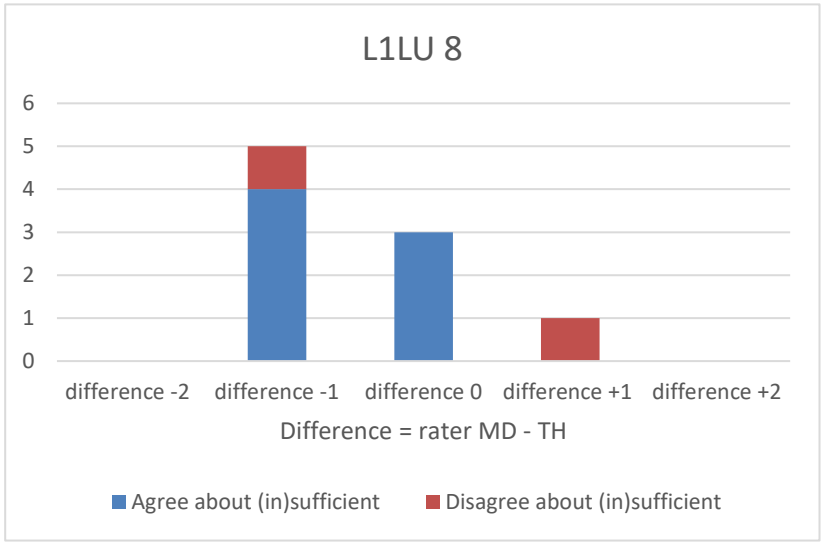
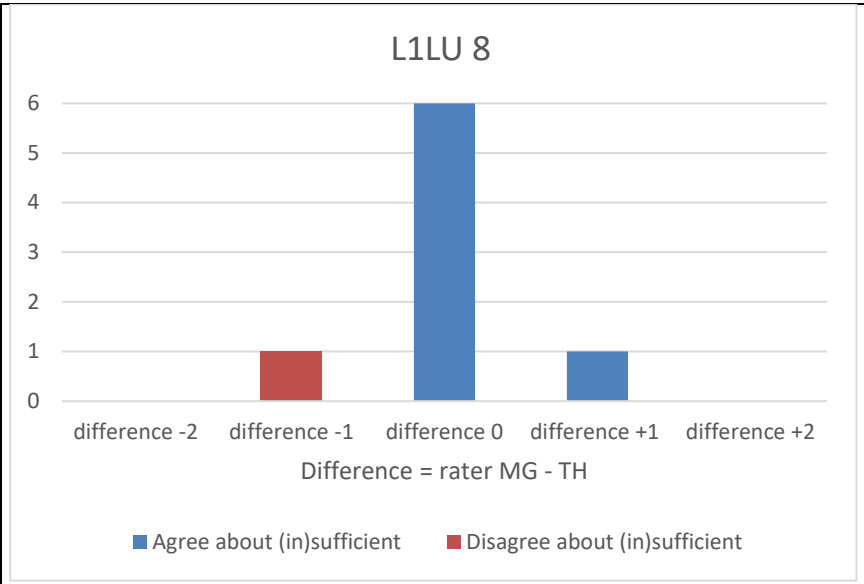
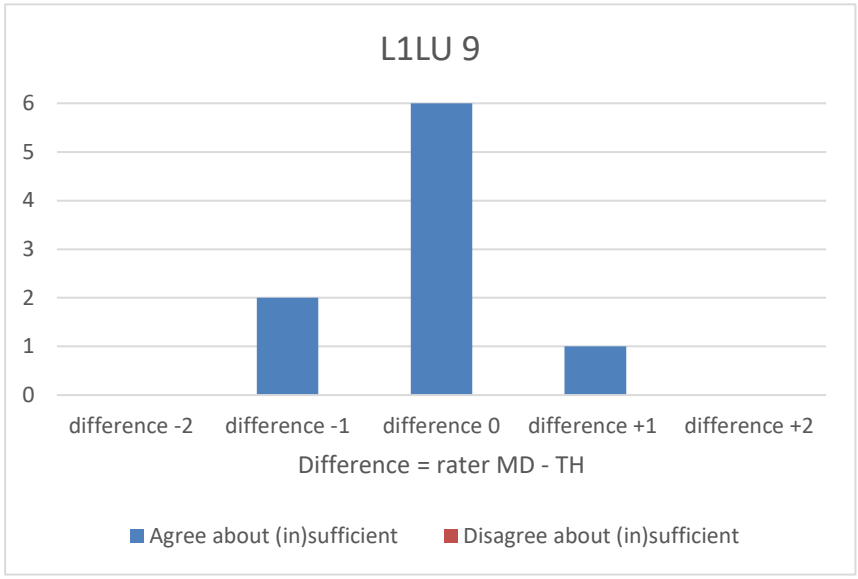
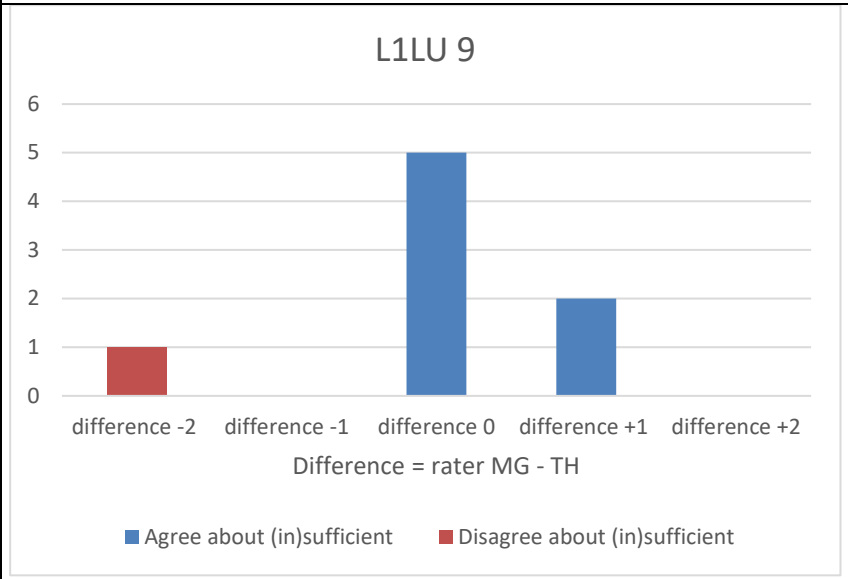
L1LU5

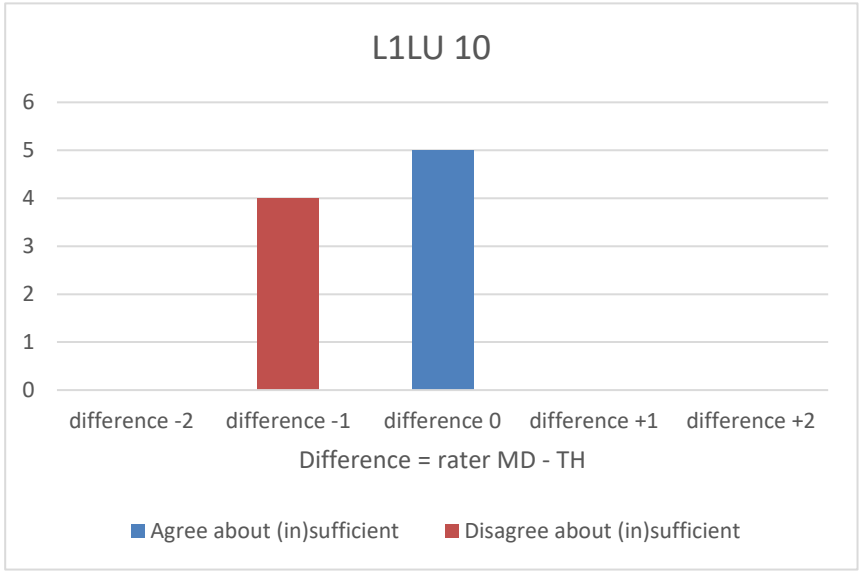
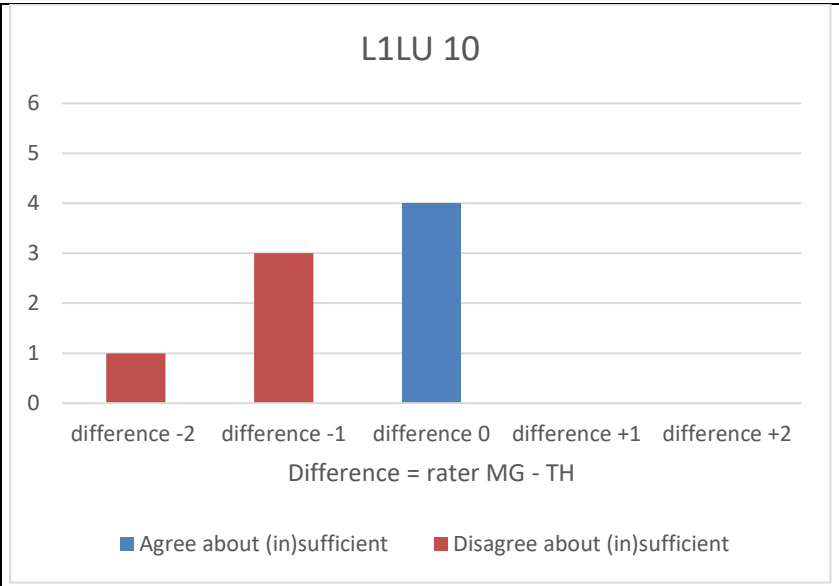
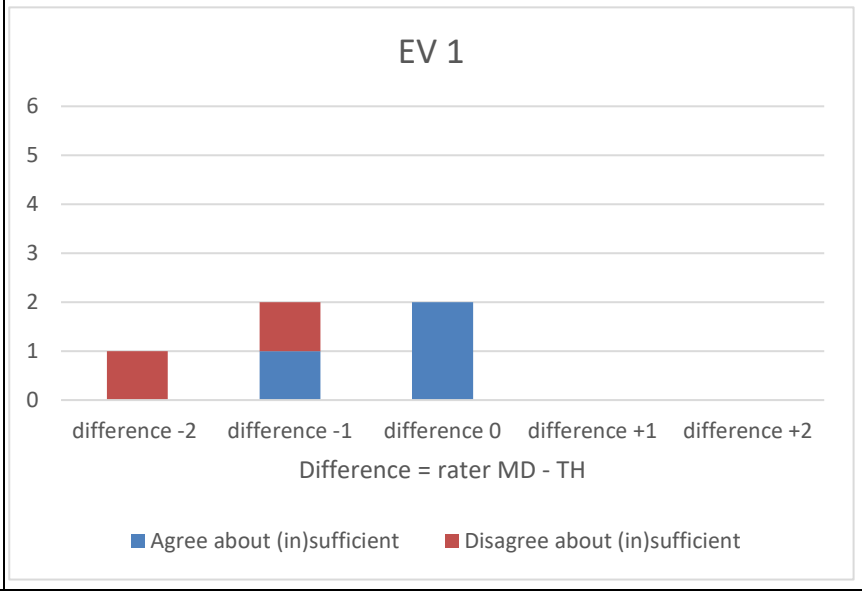
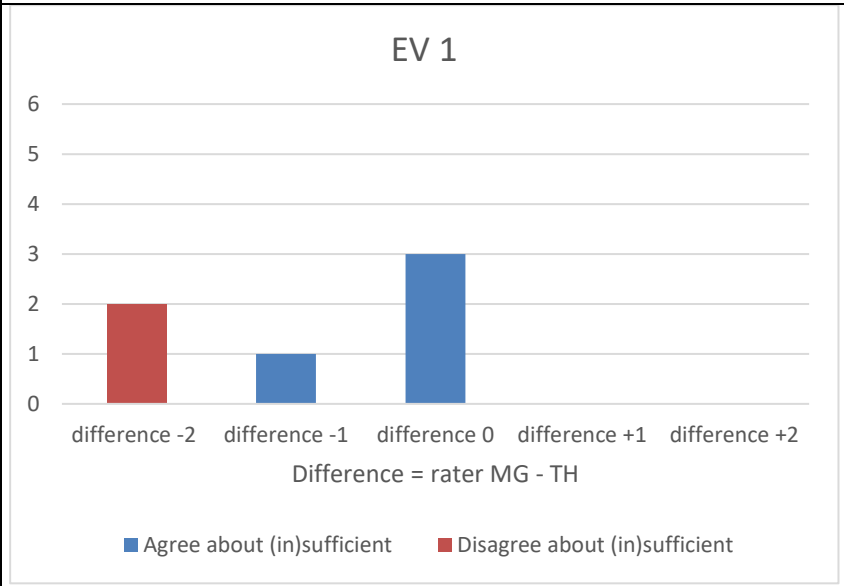


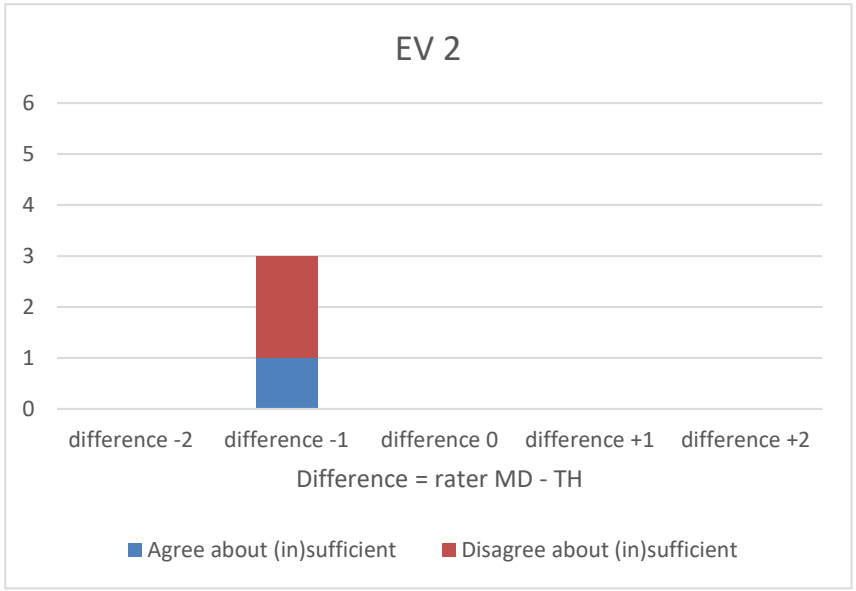
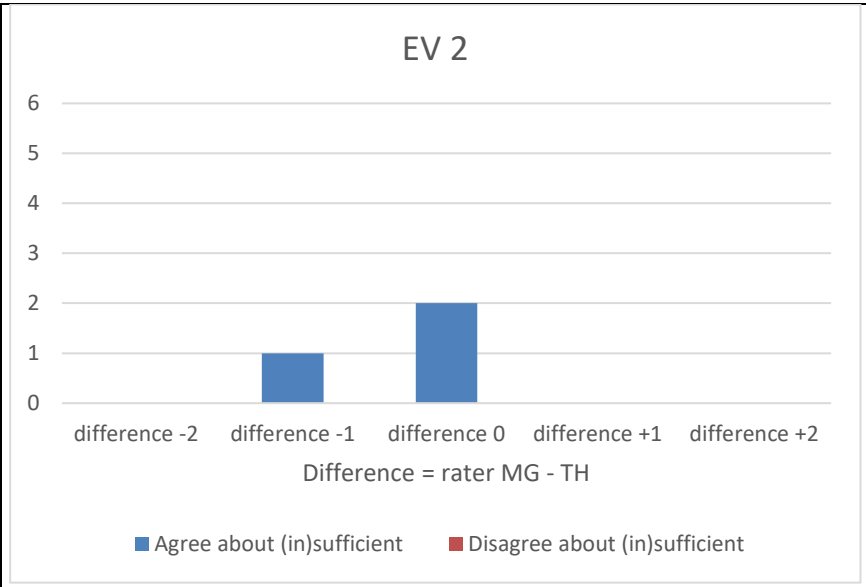
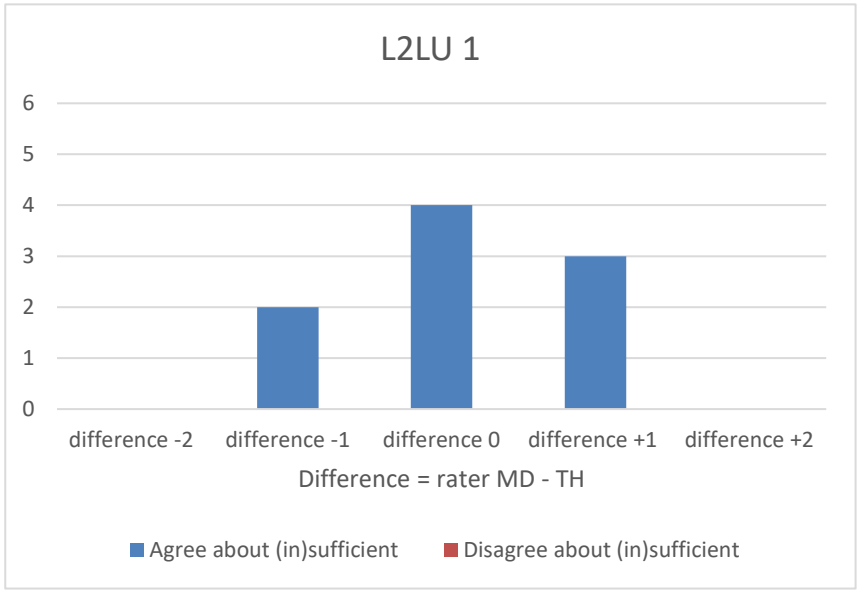
NO VARIANCE

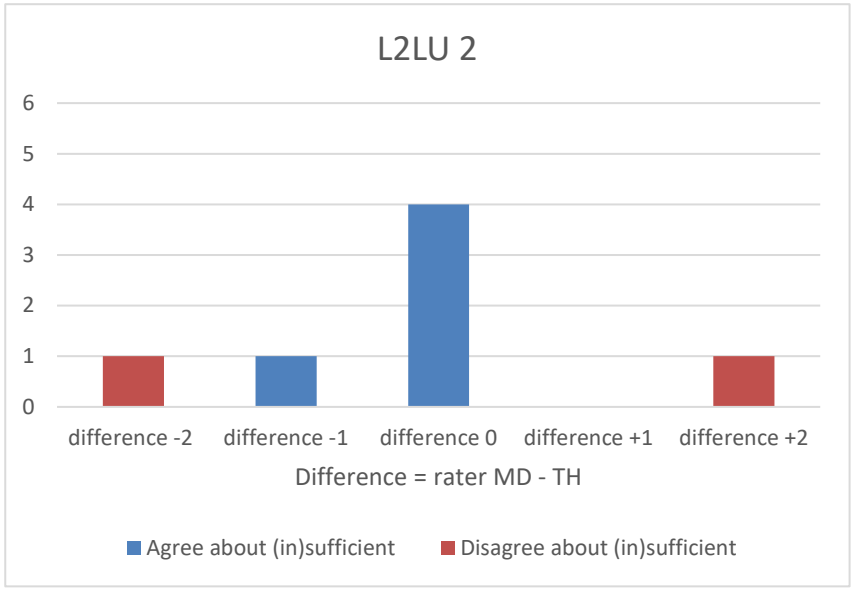
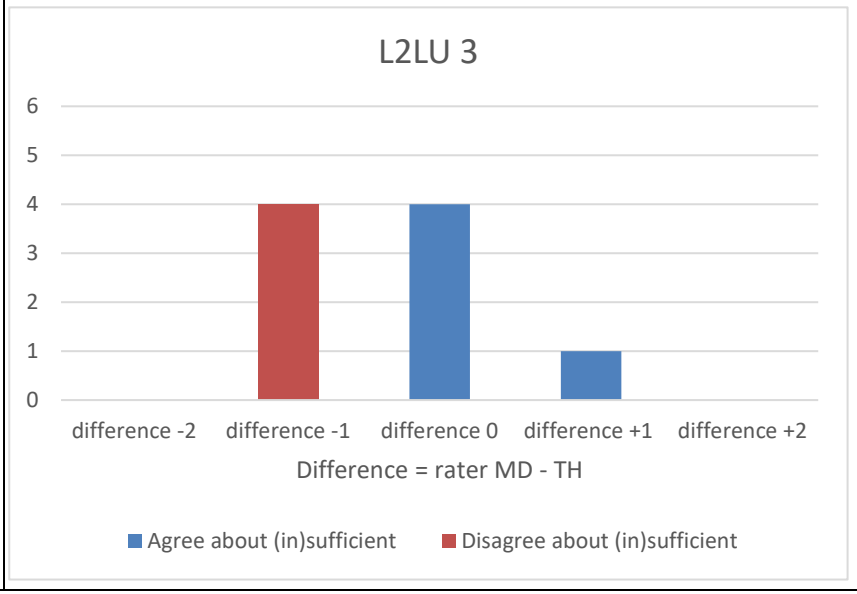


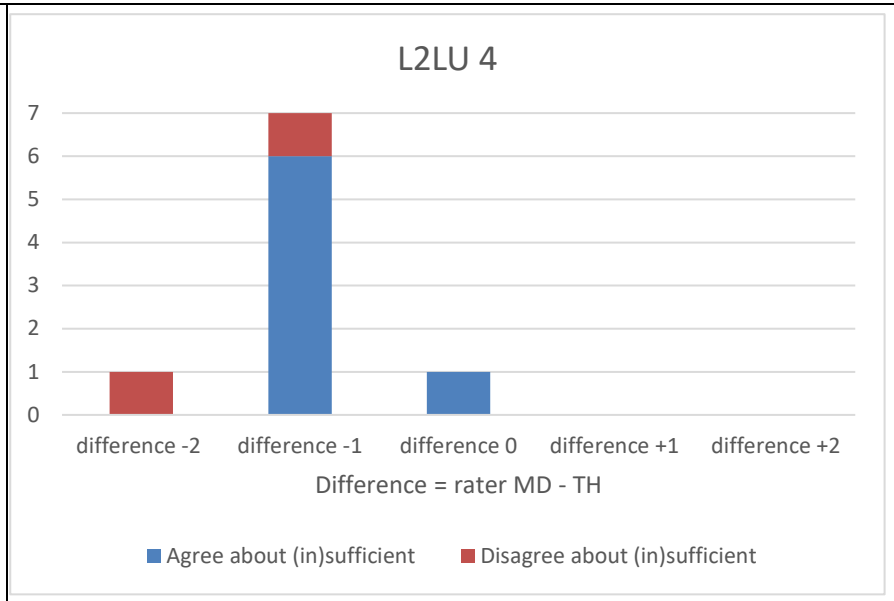
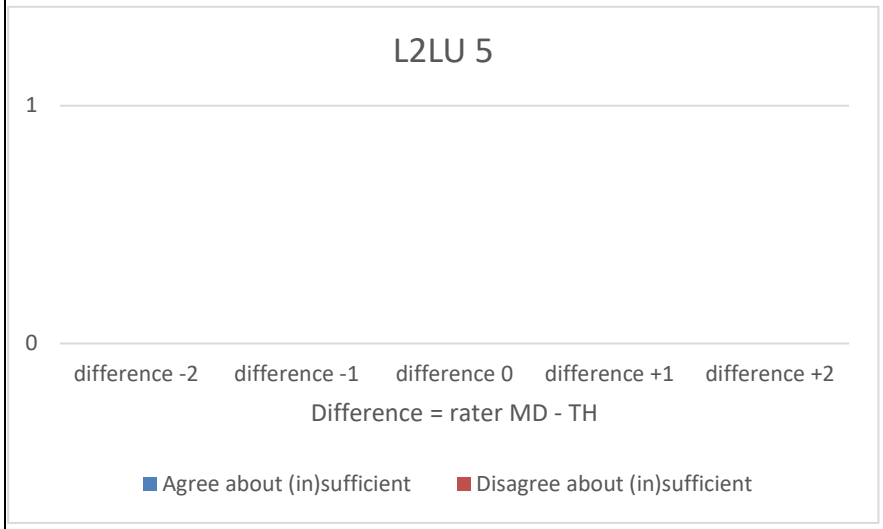
L1LU6	 L1LU 6 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 L1LU 6 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient
L1LU7	 L1LU 7 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 L1LU7 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient

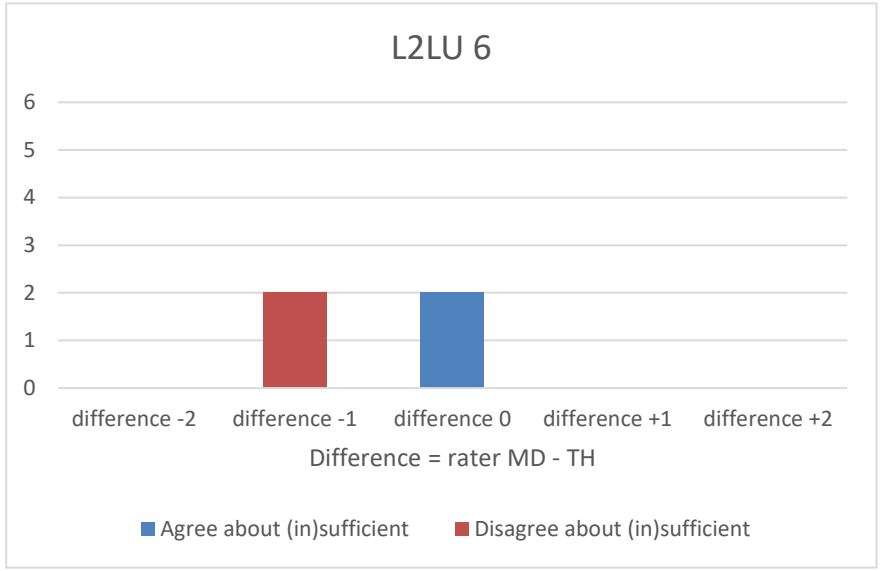
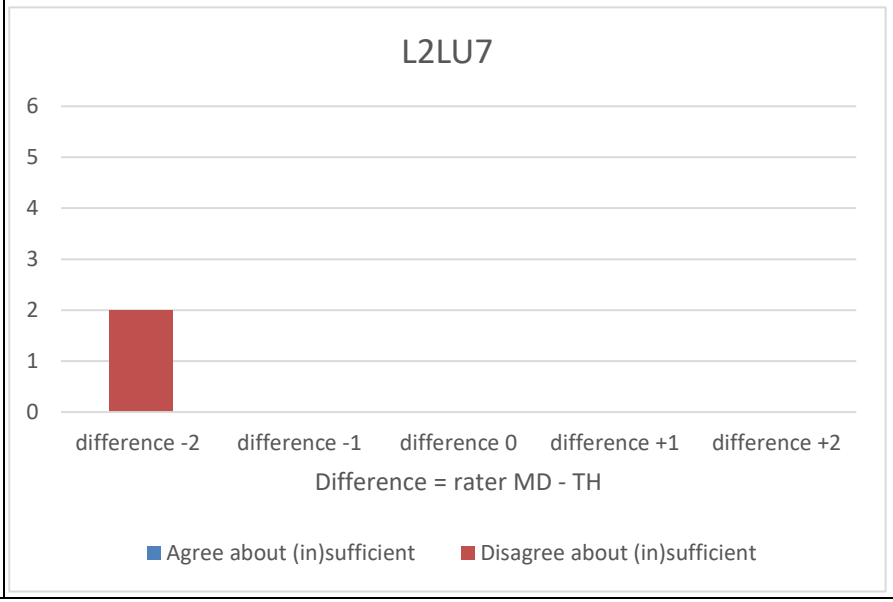
L1LU8	 L1LU 8 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 L1LU 8 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient
L1LU9	 L1LU 9 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 L1LU 9 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient

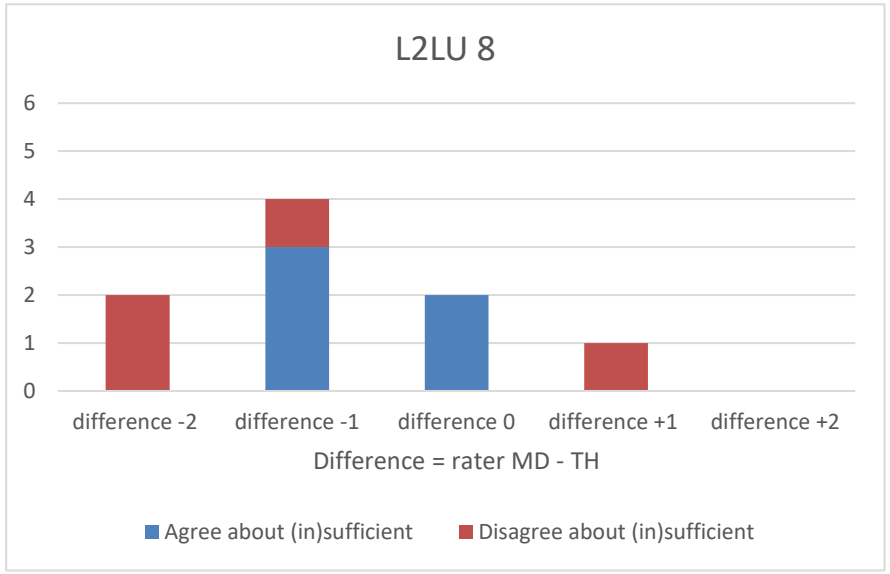
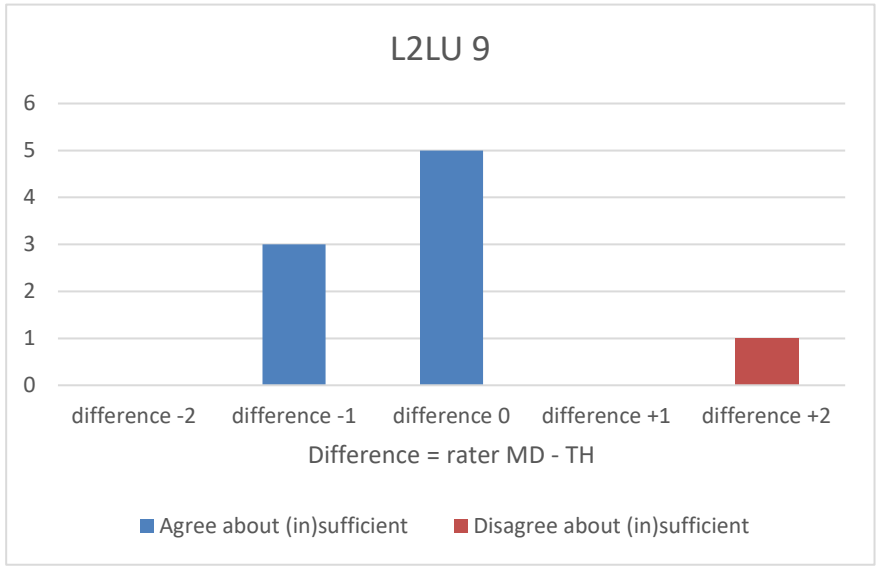
L1 LU10	 L1LU 10 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 L1LU 10 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient
EV1	 EV 1 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 EV 1 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient

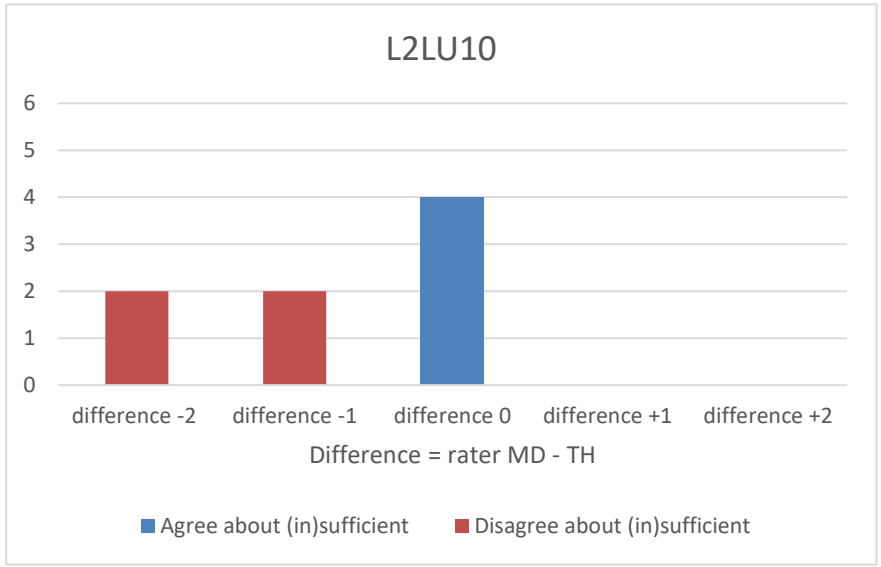
EV2	 EV 2 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	 EV 2 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MG - TH Agree about (in)sufficient Disagree about (in)sufficient
L2LU1	 L2LU 1 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	

L2LU2	 L2LU 2 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	
L2LU3	 L2LU 3 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	

L2LU4		
L2LU5	NO VARIANCE 	

L2LU6	 L2LU 6 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	
L2LU7	 L2LU7 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	

L2LU8	 L2LU 8 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	
L2LU9	 L2LU 9 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	

L2 LU10	 L2LU10 6 5 4 3 2 1 0 difference -2 difference -1 difference 0 difference +1 difference +2 Difference = rater MD - TH Agree about (in)sufficient Disagree about (in)sufficient	

Appendix I. Version 3.0 of the ADAPT-instrument

(confidential, removed for public version)