Lindhardsen, Vivian

Doctoral Thesis

From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviours

PhD Series, No. 1.2009

Provided in Cooperation with:
Copenhagen Business School (CBS)

Suggested Citation: Lindhardsen, Vivian (2009) : From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviours, PhD Series, No. 1.2009, ISBN 9788759383773, Copenhagen Business School (CBS), Frederiksberg, https://hdl.handle.net/10398/7743

This Version is available at:
https://hdl.handle.net/10419/208709

From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviors

by

Vivian Lindhardsen

Institut for Internationale Sprogstudier og Videnskabsteknologi

Copenhagen Business School

Table of Contents

Acknowledgements

My first thanks go to my two supervisors: Kirsten Haastrup and Alister Cumming for their excellent advice and support throughout my study.

My thanks also go to Karsten Sielemann for helping me apply for financial support from the Ministry of Education and for introducing me to the raters of the HHX exam. I also thank the HHX raters, the participants in my study for their thorough work.

I also thank Lise-Lotte Hjulmand and Alex Klinge for allowing me to work on my study from abroad.

Last, but not the least, I thank my husband, Michael, and my children, Marcus, Alexander, and Sophia.

Dedication

Dedicated to my husband, Michael, and my children, Marcus, Alexander, and Sophia

Abstract

The present study maps the decision-making behaviors of experienced raters in a well-established Communal Writing Assessment (CWA) context, tracing their behaviors all the way from the independent rating sessions, where the initial images and judgments are formed, to the communal rating sessions, where the final scores are assigned on the basis of collaboration between two raters. Results from think-aloud protocols, recorded discussions, retrospective reports and reported scores from 20 raters rating 15 ESL essays show that when moving from the independent ratings to the communal ratings, there is little, if any, increase in rater agreement levels and the raters' attention to the textual features corresponding to the official criteria become more evenly distributed. However, rather than consulting the scale descriptors directly in resolving insecurities about score assignment, the raters seemed to rely heavily on each others' expertise, thereby reducing the importance of the scale and emphasizing the value of the community of raters.

In validating their scores in the communal rating discussions the raters appeared to be critically and equally engaged in the discussions, and through deliberating and refining their assessments the raters believed that CWA practices produce more accurate scores than in independent ratings and lead to professional development. These interpretations support a hermeneutic rather than a psychometric approach to establishing the validity of the present CWA practices.

List of Tables

List of Figures

List of Appendices

Chapter 1
Introduction

1.1 Purpose of the Study

The purpose of the present thesis study was to describe and analyze the decision-making behaviors of raters assessing EFL (English as a Foreign Language) essays in a CWA (Communal Writing Assessment) practice, more specifically to trace the progression of their behaviors from when they form a preliminary score in independent rating sessions to when they collaborate with another rater to reach a final score in communal rating sessions. The study was stimulated by recent introspective studies into raters' decision-making behaviors in traditional performance-based writing assessment and by recent studies into CWA practices, thus continuing the increasing focus in research and in assessment practices on raters' rating process.

1.2 Rationale for the Study

The development in writing assessment towards focusing on the rating process (how raters reach their scores) rather than just the rating product (the final scores) reflects a growing interest in other aspects besides reliability in validating performance-based writing assessment. Since low inter-rater reliability rates were reported in performance-based writing assessment in the 1960s (especially Diederich, French & Carlton, 1961), major standardization procedures have been introduced to reduce rater variance in scoring. Recently, however, although inter-reliability remains a central concern, researchers and practitioners alike have begun to look into how raters reach their scores in the validation of performance-based writing.

The past couple of decades have witnessed a steady increase in studies that investigate the rating process (Connor-Linton, 1995a; Cumming, 1990; Cumming, Kantor & Powers, 2001, 2002; DeRemer, 1998; Erdosy, 2004; Huot, 1993; Lumley, 2002, 2005; Milanovic, Saville & Shuhong,

1996; Pula & Huot, 1993; Sakyi, 2000, 2003; Vaughan, 1991; Wolfe et al., 1998). These studies have refuted "the assumption that trained raters will respond to an essay in the same way if they are given a set of characteristics to guide them" (Vaughan, 1991:111), and they have shown that the rating task is a highly complex one in which raters rely on their own individual styles, invoke different assessment criteria, and at times feel frustrated having to align the simple scoring rubrics to the complex student scripts and their own responses to them (Broad, 1994, 2000; Lumley, 2002, 2005).

The increasing attention to the rating process is not restricted to research methods, but can also be seen in some assessment practices. In communal assessment practices such as CWA where "two or more raters work(ing) together to reach a joint decision on the basis of a writing performance" (Broad, 1997:134) the focus on the rating process is accentuated. Raters are here given an opportunity to uncover part of their rating process in that they can bring forward their different assessment strategies and have them validated before a final score is assigned. The validity potentials of CWA have been emphasized by an increasing number of scholars in educational measurement (e.g. Broad, 1997; Broad & Boyd, 2005; Moss, 1994, 1996; Moss & Schutz, 2001; Moss, Schutz & Collins, 1998) as well as in composition (e.g. Broad, 2003; Huot, 1996). The arguments for increased validity potentials in CWA are grounded in the paradigm of a social construction of reality (Berger & Luckmann, 1966), in which differences are not necessarily seen as a "measurement error" but are accepted as the norm and as a strength that can potentially bring about a synergy for sounder assessments and professional development (Moss, 1994, 1996).

Despite Broad's claim on CWA that "the limited application of such methods to writing assessment has been conducted by researchers, not practitioners" (Broad, 2003:14), communal rating procedures, have, in fact, been practiced for many years in Denmark at all educational levels. With the sound theoretical foundation for CWA and its application in at least some countries, it is surprising that so few empirically based studies have been conducted on such writing assessment practices. Some empirical studies have been carried out (Allen, 1995; Broad, 2000, 2003; Condon & Hamp-Lyons, 1994; Durst, Roemer & Schultz, 1994; Johnson, Penny,

Shumate, & Fisher, 2005; Mohan & Low, 1995; Moss et al., 1998; Nixon & McClay, 2007), but they are sparse and eclectic, and so the research body on CWA needs to be expanded by further exploring what actually takes place during CWA sessions.

1.3 The Present Study

The present study is an empirical investigation of the decision-making behaviors of experienced raters in a well-established CWA practice. The thesis research systematically traces the progression of raters' decision making from the independent rating sessions, where they rate student scripts individually, to the communal rating sessions, where two raters are paired to reach a joint decision on a score. The study draws on data obtained from think-aloud protocols from independent rating sessions and recordings from verbal exchanges in communal rating sessions, as well as from raters' retrospective perceptions. The background for the study is the Danish HHX (Højere Handelsskole Eksamen) exam, a high stakes business-focused high-school exit exam, and the participants (the raters) in this study are all members of the HHX censorkorps (national rater corps) for the HHX written EFL component. The scripts are essays written either by students who have sat for the HHX exam or by students who have practiced for this exam.

The study relates to and expands the existing research into raters' decision-making behaviors in writing assessment. It supplements studies into the rating process in independent ratings and along with a limited set of researchers (e.g. Allen, 1995; Broad, 2000, 2003; Mohan & Low, 1995; Nixon & McClay, 2007) challenges the dearth of empirical research into CWA. What makes this study unique is that it systematically traces raters' decision-making all the way from independent rating sessions to communal rating sessions and documents the difference in rating behaviors between the two sessions. Further, it records the behaviors of raters highly experienced in rating in a well-established CWA practice.

In shedding light on the complexities of decision-making behaviors in a well-established CWA system, this study intends to contribute to the validation research of CWA.

1.4 Organization of the Thesis

The thesis is organized around the empirical study of raters' decision-making behaviors of a well-established CWA practice and the research which has stimulated this study. Chapter 2: Literature Review is divided into three parts: Independent Ratings, Communal Ratings, and From Independent Ratings to Communal Ratings. Part 1, Independent Ratings, reviews the literature on raters' decision-making behaviors in traditional, independent ratings of performance-based writing samples, focusing on studies which have employed think-aloud protocols to map the decision-making process. The results of these reviewed studies are analyzed in terms of their validity implications of traditional, independent ratings of written scripts. Part 2, Communal Ratings, continues the focus on raters and reviews the sparse literature on rater dynamics in CWA. The studies are reviewed in terms of the validity potentials of CWA. Part 3, From Independent Ratings to Communal Ratings, describes briefly how the implications from the reviewed literature in Independent Ratings and Communal Ratings have generated an interest in mapping raters' decision-making behaviors in a well-established CWA practice focusing on the development of these behaviors from when the raters form their preliminary judgments of student scripts in independent rating sessions to when they reach a final score on these scripts with a co-rater in communal rating sessions. Chapter 3: Methods describes the venue of the empirical study, a well-established CWA practice in Denmark, and how raters' decision-making behaviors were chronicled with the use of verbal reports: think-aloud protocols in independent rating sessions, recording the rater discussions in communal rating sessions, and raters' retrospective reports on their CWA experiences. The results are presented in three chapters. Chapter 4, which maps the raters' decision-making behaviors in their independent rating sessions; Chapter 5, which maps the raters' behaviors in their communal rating sessions; and finally Chapter 6, which compares the results from chapters 5 and 6 to trace the progression of the raters' decision-making behaviors from the independent rating sessions to the communal rating sessions. Chapter 7: Discussion and Conclusion sums up the findings and discusses them in terms of validity implications of CWA and the concept of rater expertise.

Chapter 2
Literature Review: Raters' Decision-Making Behaviors in Writing Assessment

2.1 Purpose and Scope of the Literature Review

The purpose of this chapter is to review the literature that stimulated the present thesis study into raters' decision-making behaviors in Communal Writing Assessment (CWA). It begins with a brief account of the increasing attention to the rating process in performance-based writing assessment. The rest of the chapter is divided into three parts (Independent Ratings, Communal Ratings, and From Independent Ratings to Communal Ratings), each part emphasizing the increased focus on raters' rating process in writing assessment as it is manifested in research as well as in actual assessment practices. The first part, Independent Ratings, reviews the expanding and insightful literature into raters' decision-making in traditional performance-based writing assessment practices, in which raters rate student scripts independently. The results of these mainly introspective studies present the rating task as a complex and multi-faceted one, in which raters face the difficult task of reconciling their complex readings of student scripts with simple and abstract scoring rubrics. The results further show that although prototypical rating behaviors can be identified, raters exhibit great variability in how they approach the rating task. The results of the reviewed literature are analyzed in terms of validity implications of traditional, standardized writing assessment practices. The second part, Communal Ratings, reviews the sparse and less structured literature on CWA, a communal assessment practice in which raters first rate student scripts independently and subsequently meet with (an)other rater(s) in a communal rating session to collaborate on a final score assignment. It presents the growing theoretical interest in CWA, and communal assessment practices in general, as lying within a hermeneutic paradigm rather than a psychometric paradigm. It discusses the theoretically based validity claims on CWA and reviews the sparse and exploratory empirical research into such assessment practices. The third part, From Independent Ratings to Communal Ratings presents the purpose of the present thesis study, which is to map raters' decision-making behaviors in

CWA by tracing their progress all the way from their independent rating sessions to their communal rating sessions. It thus continues the focus on the rating process by building on research into raters' decision-making behaviors in traditional, independent writing assessment practices and the sparse research into CWA practices.

2.2 Increasing Attention to the Rating Process in Performance-Based Writing Assessment

Performance-based writing tests such as essays, reports, reviews, etc. are now a recognized and widely used instrument to test students' writing abilities (Barkaoui, 2007a)[1]. Unlike more indirect tests such as multiple choice tests, this form of assessment requires the students to integrate multiple skills and knowledge in constructing their responses and allows them some latitude in responding to their task (Weigle, 2002). The complex student responses that such tests stimulate inevitably lead to variation not only in the student responses themselves (e.g. Barkaoui, 2007b: Breland, Lee, Najaran & Muraki, 2004; Lee, Breland, & Muraki, 2004; Polio & Glew, 1996), but also in the interpretation and judgment of these responses (e.g. Barkaoui, 2007b; Broad, 2003; Huot, 2002; Lumley, 2005). In particular, Diederich et al.'s (1961) study reported an alarmingly high variance in rater agreement (median correlation of .31 between raters) in performance-based writing assessment.

Since the publication of Diederich et al.'s study (1961) rigorous reliability-boosting standardization procedures have been implemented that reduce the human "measurement error" of variation considerably. By now most large-scale writing assessment instruments come with detailed scoring rubrics with scale descriptors and benchmark samples accompanied by elaborate rater training and rater monitoring programs that train raters to use rubrics consistently and uniformly. Research into such procedures concludes that high inter-rater reliability levels can be obtained if raters read scripts quickly and superficially while sticking closely to the scoring rubrics (e.g. Myers, 1980; Charney 1984; Stansfield & Ross, 1988; Weigle, 1994), but also that

[1] Spolsky (1995) recounts in detail how performance-based assessment fell out of popularity for some time, especially in the US.

idiosyncratic differences cannot be eliminated (Kondo-Brown, 2002, Sweedler-Brown, 1985). The pressure on raters to rate in a standardized way is often high, as their suitability as professional raters often depends on a high reliability level. In other words, a proficient rater is a reliable rater, who can rate fast and without personal commitment (as illustrated by Wolfe, Kao & Ranney, 1998).

Although focus on inter-rater reliability has not waned, a growing number of researchers and practitioners have in the past couple of decades begun to question the notion that reliability is the sole method or criterion in validating scoring procedures in performance-based writing assessment and have argued that validity must remain the essential concern in validating a test. Williamson even challenged the traditional notion that reliability is a precondition for validity, arguing that, "comparatively high reliability is neither a necessary nor a sufficient condition for establishing the validity of a measure" (1994:162). Some scholars entirely dismiss the notion of reliability in the traditional psychometric sense or go so far as to argue that an excessive focus on reliability and its accompanying privileging of standardization undermine or even corrupt the validity of a measurement. Pamela Moss, a prominent theorist in educational assessment, has been particularly critical of the psychometric notion of reliability. In asking the title question "Can there be validity without reliability? her answer is a convincing "yes", if by reliability is meant "consistency, quantitatively defined, among independent observations or sets of observations that are intended as interchangeable" (Moss, 1994:6).

The last couple of decades have witnessed an expanding attention to other aspects besides reliability in validating performance-based writing, and it is now acknowledged that variability is inevitable in writing as well as reading (Barkaoui, 2007b; Broad, 2003; Deville & Chalhoub-Deville, 2006; Huot, 2002). In scoring, this has manifested itself in valuable research into how raters reach their judgments during their rating process, and it has inspired assessment practices such as CWA that allows raters to validate their assessments against other raters during the rating process. Section 2.3: Independent Ratings below reviews research into the rating process in traditional performance based writing assessment, and Section 2.4: Communal Ratings reviews literature on CWA.

2.3 Independent Ratings

Research into what goes on "behind the curtain" (Connor-Linton, 1995b) of raters' minds when they assess written scripts independently has blossomed over the past couple of decades. This inquiry has resulted in progressively detailed accounts of not only what raters attend to while rating their scripts, but also how they conduct themselves during the rating process.

2.3.1 Simple Accounts of Raters' Decision-Making Behaviors

Early accounts of raters' decision-making behaviors were relatively simple in nature: They employed indirect methodologies to account for what textual features affected raters' decision-making and presented simple models of the rating process.

2.3.1.1 Textual Features

Many studies have examined written scripts for traits (t-units, grammatical errors, vocabulary, length, spelling, etc.) associated with high or low scores on writing tests. Such studies have produced varying results. High scores in EMT (English Mother Tongue) tests have been associated with vocabulary density (Grobe, 1981), with length and freedom from errors (Stewart & Grobe, 1979), and with quality of content (Freedman, 1979). In ESL (English as Second Language) language-related features, especially the absence of error, have been shown to contribute significantly to the scores of ESL scripts (e.g. Homburg, 1984; Song & Caruso, 1996; Sweedler-Brown, 1993).

2.1.1.2 Rating Process

One of the first researchers to suggest a model of the rating process was Homburg (1984). He proposed the often cited "funnel model" in which it is suggested that raters grossly categorize student scripts on the basis of one textual feature and then categorize further on the basis of yet

other features. Although his model offers some information on what features seem to be significant in the judgment of student scripts, it has been criticized for being much too simple in that it suggests a linear process of decision-making, based entirely on countable textual features of the scripts.

Another widely cited model of raters' decision-making that surfaced around the same time as Homburg's funnel model was Freedman and Calfee's (1983) model of raters' rating process. It distinguished itself from Homburg's (1984) model by being grounded in cognitive models of information processing. Freedman and Calfee suggested that raters go through three main stages during the course of their rating. First, raters read and comprehend the written script to create an image of it, then they evaluate this constructed image, and finally they articulate a judgment based on a comparison of the text image and the scoring rubric. Although they assumed that raters will have to have created an image of the student script before evaluating that image, Freedman and Calfee acknowledged that this process is often recursive rather than strictly linear in that the raters may not finalize their interpretations of the entire script before judging it. Besides suggesting recursion, their model allows for variation. Variation is inevitable as raters may draw on different knowledge sources, beliefs, and value systems when interpreting the scripts as well as when evaluating and judging them. Later studies have confirmed this variability by documenting that raters consult different sources to form their assessments of student scripts. Thus, Pula and Huot (1993) found that raters' prior knowledge—especially personal background, previous professional training and work experience—impacts on raters' rating behavior. Erdosy (2004) confirmed this variability, stressing raters' differences in personal backgrounds and professional experiences. Differences in culture and disciplinary background have also been found to influence rater behavior (e.g. Connor-Linton, 1995a; Mendelsohn & Cumming, 1987; Shi, 2001).

Freedman and Calfee's (1983) information processing model of rating behavior has contributed to the understanding of raters' overall decision-making, especially the notion that raters do not evaluate the student scripts, but rather their constructed image of them. But their model has been criticized for being far too simple. It does not allow for the complex, multifaceted, and

interactive nature of decision-making identified in later studies (e.g. Cumming, 1990; Cumming et al. 2001, 2002; DeRemer, 1998; Lumley, 2002, 2005; Milanovic et al. 1996; Sakyi, 2000, 2003; Wolfe et al., 1998). Moreover, the model takes into consideration only variation in the textual features that may impact on raters' interpretations, not possible variation in the raters' cognitive processing actions (Cumming, 1990; Cumming et al. 2001, 2002; Lumley, 2002, 2005; Sakyi, 2000, 2003; Vaughan, 1991; Wolfe et al. 1998).

2.3.2 Complex Models of Raters' Decision-Making Behaviors

Recent studies into raters' decision-making behaviors have presented detailed accounts of what raters do during the course of rating students' scripts. Employing a more direct and process-oriented methodology, such as think-aloud procedures (TA), these studies have tended to focus on raters' comments, making it possible to focus on the raters' own images that they have constructed of the student scripts, not the scripts themselves. These studies have contributed to information about what textual and contextual features raters attend to during their rating process, and they have shed light on how raters direct and monitor their rating behavior during the course of rating. The picture that has emerged of raters' decision-making from this inquiry is that raters engage in complex, multi-faceted rating behaviors.

2.3.2.1 Interpretation Strategies and Judgment Strategies

Grounded in Freedman and Calfee's (1983) premise that raters construct an image of student scripts and subsequently evaluate and judge that image, Cumming (1990), and later Cumming et al. (2001, 2002), identified a multifaceted set of decision-making behaviors involved in rating writing for ESL/EFL tests. The set comprised 27[2] behaviors (reduced from 35 in Cumming et al., 2001, 2002), divided into interpretation strategies and judgment strategies. Interpretation strategies are strategies raters use to create an image of the student scripts; judgment strategies are used to evaluate and judge that image. Cutting across the interpretation strategies and the judgment strategies, Cumming (1990, 2001, 2002) identified textual and non-textual focus areas.

[2] In Cumming's (1990) earlier study, which was based on a smaller body of data, 28 decision-making behaviors were identified.

The former showed how raters focus on the scripts (or, rather, their image of the scripts) and the latter how they manage or control their own decision-making. The broad categories of their decision-making matrix are reproduced in Figure 2.1.

	Self-Monitoring Focus	Rhetorical and Ideational Focus	Language Focus
Interpretation Strategies			
Judgment Strategies			

Figure 2.1: Cumming et al.'s Framework of Raters' Decision-Making Behaviors

The matrix identifies what the raters focus on to construct an image of student scripts (interpretation strategies) and what they focus on while evaluating or judging the scripts (judgment strategies).

During the process of developing an image of the scripts, the raters in Cumming's studies were shown to focus on the scripts by discerning the rhetorical structure, summarizing ideas, classifying errors, and interpreting or editing ambiguous phrases; and to focus on their self-monitoring behavior by engaging in such behaviors as reading the scripts and envisioning the personal situation of the writer. When judging their image of the scripts, the raters would focus on the scripts by assessing logic or topic development, task completion or relevance, coherence, originality or creativity, organization, style, use and understanding of source material, ideas and rhetoric, total written production, comprehensibility and fluency, frequency and gravity of errors, lexis, syntax or morphology, spelling or punctuation, and language overall.

Sakyi (2003), too, made a distinction between interpretation strategies and judgment strategies, although not as explicitly as Cumming (1990) and Cumming, et al. (2001, 2002) in that he distinguished between raters' understanding the text and identifying or correcting errors on the one hand and evaluating content as well as structure and format on the other.

Although the decision-making behaviors in Cumming et al.'s (2001, 2002) model are logically distinguishable, the authors acknowledge that they are interrelated and frequently occur in conjunction with one another. As a consequence of this interrelatedness, the rhetorical focus area and the ideational focus area of Cumming (1990) were consolidated in Cumming et al. (2001, 2002). Although the language and content focus areas are also to some extent interrelated, Cumming (1990) found that raters attempted to distinguish between language proficiency on the one hand and content and ideas on the other.

While Cumming's (1990) and Cumming's (2001, 2002) studies confirm Freedman and Calfee's (1983) model that raters construct a textual image (by using interpretation strategies) and evaluate and judge that image (by using judgment strategies), the data in Cumming (1990) and Cumming (2001, 2002) revealed a more complex, interactive, and multifaceted rating process than that suggested by Freedman and Calfee (1983).

Other researchers have also used direct research methodologies to consult raters' about their comments directly. They have drawn on Freedman's and Calfee's (1983) model to deepen our understanding of raters' complex rating behaviors. Although they have not maintained as strict a distinction between interpretation strategies and judgment strategies as Cumming (1990) and Cumming et al. (2001, 2002) and have not classified or termed rater focus in exactly the same way, similar studies into the rating process have confirmed raters' focus on various textual features and self-monitoring aspects, and they have contributed to understanding how raters sequence their behaviors and how they use a scoring rubric in assigning their scores (e.g. DeRemer, 1998; Huot, 1993; Lumley, 2002, 2005; Milanovic et al. 1996; Sakyi, 2000, 2003; Vaughan, 1991; Wolfe, 1997; Wolfe et al., 1998).

2.3.2.2 Textual Focus

Many of the above mentioned studies have detected patterns of how raters balance their attention to the textual features of the scripts. The weight that raters attribute to different text features in their assessment decisions has been of particular interest.

Although it is not possible to unequivocally measure the weight that raters attach to the different textual features they encounter in the student scripts, a number of researchers have attempted to identify certain tendencies. This has often been done by counting raters' comments on the assumption that the frequency of mention corresponds to the weight attached to different textual features. Counting the frequency of raters' comments may give some indication of the importance raters attach to different textual features, but that is an uncertain assumption. As Lumley pointed out "the value of relative importance or influence of comments made under different categories cannot be evaluated. Thus a single comment under task fulfillment or coherence may well carry more weight in the rater's judgment than several comments on individual spelling or tense errors" (2005:193). Counting comments by raters, however, has been the preferred method of getting at the raters' priorities when it comes to discerning features of students' scripts and their contributions toward a final score. Admitting that it is impossible to establish the exact weight that raters attribute to different textual features, Milanovic et al. (1996) nevertheless concluded on the basis of their empirical study that "the relatively modest number of characteristics catalogued by the markers in their written and verbal reports on each paper suggests that they only remarked upon the elements which distinguished a paper" (1996:100).

Studies that have counted raters' comments have shown that raters attend to a variety of textual features related to language, but also to content, style, length and rhetoric. Vaughan (1991) placed raters' comments in order of frequency and reported the following order of textual weight: content, handwriting, tense/verb problem, punctuation, introduction, morphology[3]. Although not presenting raters' comments by frequency, Milanovic et al. (1996) recorded raters' priority of comments on the following features: length, legibility, grammar, structure, communicative effectiveness, tone, vocabulary, spelling, content, task realization, and punctuation. Addressing slightly more specific areas of content and organization, Sakyi (2000 and 2003) recorded most raters' comments on the following features: introduction, thesis

[3] That content was placed highest on the list does not necessarily show that the broad category of content carried more weight than the broad category of language, as language here was split into subcategories (e.g. tense/verb problem and morphology), each competing against the broader category of content. The list does show, however, that content was referred to as a broad category, that language was referred to more specifically, or that content as a broad category received the most comments.

development, topic relevance, organization, supporting argument, intelligent thinking, grammar, vocabulary, sentence structure, grammatical errors. These and other studies into raters' decision-making (e.g. Cumming, 1990; Cumming et al. 2001, 2002; Lumley, 2005) have shown that raters make reference to a wide spectrum of textual features.

Although variability has been found in the importance that raters attach to different textual features, a number of studies have noticed that raters pay particular attention to treating errors. Cumming et al. (2001, 2002) and Cumming (1990) reported that raters devoted much of their attention to error treatment. The rating behaviors that stood out in their study were editing phrases and classifying language errors. All other behaviors accounted for less than 10% of the total data. Sakyi (2003), too, observed that errors took up much of raters' attention. In particular, Sakyi found that if raters did not correct errors, at least they couldn't help but identify them (in mechanical, grammatical and syntactic categories). This focus on errors supports earlier research using indirect methodologies (referred to above in Section 2.3.1.1) to measure raters' assessment criteria. Lumley (2005:186), referring to other studies beside his own, reports that command of grammar seems to attract a lot of raters' attention. Huot (1993), who studied EMT, rather than ESL writing assessment, suggests that raters may focus on errors in grammar and mechanics because they are easy to recognize.

2.3.2.3 Use of Scoring Rubric

In contrast to studies like those of Cumming and colleagues, a few researchers have investigated how raters use a designated scoring rubric (e.g. DeRemer, 1998; Lumley, 2002; 2005: Sakyi, 2000, 2003; Vaughan, 1991; Wolfe et al. 1998). Research into how raters use the scoring rubric presented to them is very important because the rubric is an explicit statement of the theoretical construct of writing ability (or at least a reflection of the test developers' view of what is of relevance in assessing writing). Most tests assume that "given a scale that describes the characteristics of an essay at each level, trained raters will assess the essays in the same way every time" (Vaughan, 1991:112).

Lumley (2002, 2005), however, concluded in his study on raters' uses of a scoring rubric that it was difficult to identify a full picture of how raters actually use a rubric as his raters tended to not articulate their use of it in determining their judgments of the scripts. Lumley's raters tended to form intuitive impressions and only make explicit references to the scale of the rubric when justifying or articulating their scores. Therefore, "the relationship between the scale contents and text quality remains obscure" Lumley (2002:246). He found that "the movement is from their own impressions towards descriptions of texts in terms of the scale descriptors, rather than from the scale towards their own 'styles' or criteria." (Lumley: 2005:293). The weak position of the rubric in the overall process of evaluating writing is echoed in other studies into how rubrics are used in the rating process.

Sakyi (2003:125) identified a range of behaviors describing how raters deal with a scoring rubric, ranging from sticking very closely to it, to relying almost entirely on their own individual criteria of what constitutes good writing. He identified the following broad, rubric-consulting behaviors:

1. Matching essay characteristics to scale descriptors;
2. Using scale level descriptors to confirm or justify a score, that is, the raters first make an impression of the scripts and subsequently use the scale descriptors to narrow down or confirm their impressions (similar to Lumley's, 2002, 2005, main findings); and
3. Using one's own set of criteria with little or no reference to the rubric.

DeRemer (1998) identified a similar continuum of uses of a scoring rubric:

1. Rubric-based scoring: raters make an extensive search of the rubric in order to match the student script with the rubric;
2. Text-based scoring or a complex recognition task: raters conduct an analysis of the scoring criteria before the score is assigned, but no search in the rubric is made during the scoring process; the text, not the rubric is the focus during the rating process (similar to Lumley's, 2002, 2005, main findings); and
3. General impression scoring: the rubric is not consulted; the score is determined based on the rater's general impression.

This broad range of rubric-consulting behaviors reported across different raters and across different scripts confirms Vaughan's (1991) and Weigle's (1994) findings that despite similar training raters do not uniformly apply a predetermined rubric. Although many raters made a conscious attempt to follow the scoring guide, Sakyi (2003) found in his research that they often fell back on their individual strategies to assign a score. In such cases, such strategies were used to arrive at a final score as comparing essays, depending on a few aspects of the scripts, re-reading to confirm a score, deducting points for certain deficiencies, and even assessing features not even cited in the scoring guide (Vaughan, 1991).

2.3.2.4 The Simplicity of the Scale and the Complexity of the Scoring Task

As Lumley (2002, 2005) has emphasized, the reason that raters often fail to match student scripts to scoring rubrics is that the rubrics fall short of representing the complex nature of the scripts and the raters' perceptions of them. Raters are usually conscientious and often wrestle to fit their intuitive impressions within a scoring scale, and even at times sacrifice their impressions of the script if they are not represented in the scale (Lumley, 2005:313). Lumley found that raters do not perceive the categories in the rubric to be discrete and that the boundaries that the raters draw between textual features are often not based on the same as those in the rubric, a finding that underpins his claim that the rubrics do not actually represent student scripts (Lumley, 2005:218).

Lumley refutes possible arguments that the answer to the problem would be to improve rating scales to better represent the complexity of written scripts and the assessments of them, arguing that scales would necessarily have to entail simplification and abstraction (2002:263):"It cannot represent the individual perspective of each rater, nor describe adequately each text. Therefore raters have to fit their own perceptions to the given procedure" (Lumley: 2005:240). The simplicity and inadequacy of rating scales and sample papers to represent the scales were also noted by Elbow, who stated that "rubrics fail to fit many papers" (1993:192). This tendency was further illustrated in Broad's (2000) study, which revealed that finding sample scripts to exemplify scoring rubrics' descriptors was a difficult and frustrating job.

The above-mentioned studies convey a highly complex picture of the rating task: It is rarely a simple exercise of matching scripts to scale descriptors. Rather, raters have to interpret and judge student scripts, and they have to abstract guidelines from the rubrics to reconcile the complexity of the scripts with the simplicity of the scale descriptors. DeRemer (1998) called this an ill-structured task because (a) it is difficult to determine whether the goal has been reached, (b) the information needed to solve the problem is not entirely contained in the problem instructions, and the relevant information to solve the task is vague, and (c) there is no "legal generator" for finding all alternatives (DeRemer, 1998:13). This highly demanding endeavor often triggers a tension (Huot, 1990) between, on the one hand, the authenticity of the raters' reader response (the raters' intuitive impressions of the scripts) and, on the other hand, the obligation to reliably match the scripts to a scale (squeezing their personal response into the institutional measure).

2.3.2.5 Sequence

Having identified the various distinct decision-making behaviors, many researchers into the rating process have attempted to identify the sequence of these behaviors.

A broad sequence of the rating process was touched upon above when referring to Lumley's (2002, 2005) findings of how raters use a scoring rubric. To recapitulate, raters typically formed their individual and intuitive impressions first and then articulated and justified these impressions through the scoring rubric. Although Lumley (2002, 2005) identified this process as essentially linear, he did acknowledge that for some raters this process was cyclical with interruptions and recursions. He identified the following broad stages in the rating process:

1. The pre-scoring stage in which a first reading of the scripts is conducted. Although the quality of the script is assessed, scores are typically not considered;
2. The scoring stage in which scores are given and justified in relation to the scoring categories (analytic scale); and
3. A finalizing stage where scores are explicitly or implicitly confirmed or revised (Lumley: 2005:310).

Other studies have been more specific in identifying the sequence of distinct decision-making behaviors during the rating process. With an emphasis of the interplay of interpretation strategies and judgment strategies Cumming et al. (2001, 2002) identified the following sequence of behaviors, showing how "raters balanced processes of interpretation with processes of judgment while exercising diverse self-control strategies and attending to numerous aspects or rhetoric and ideas and language use" (2002:88):

1. Composition scanned for surface level identification (length, format, paragraphing, script);
2. Interpretation of script along with varying degrees of intervening judgment strategies (a. classifying and assessing error types; b. identifying comprehensibility leading to assessment of language use and rhetoric; c. interpreting rhetorical strategies leading to assessment of content and organization; d. envisioning the situation and personal viewpoint of the writer); and then
3. Articulating the scoring decision (while summarizing and reinterpreting judgments)[4].

Sakyi (2003) and Milanovic et al. (1996) identified a similar process, although they took into account the internalization of the rubric. Sakyi (2003:149) described the following sequence:

1. Prescoring stage, in which the scale level descriptors are read and internalized;
2. Reading or scanning the text (portions or entire text). This is similar to Cumming et al.'s (2002:21) "scanning the composition for surface level identification";
3. Read and score composition. A rater establishes a possible range of scores based on initial impressions, leading to a proposed score after reading and summarizing and comparing to rubric. This is similar to Cumming et al.'s (2001) raters' use of interpretation strategies; and
4. Confirmation-articulation –justification stage: The rater confirms with additional evidence, e.g., descriptors from a rubric, comparisons with other essays, emphasizing weak and strong points. The score is articulated, and sometimes justified, explained or revised. Finally, the score is completed with reaffirmation. This is similar to Cumming et

[4] See Cumming et al. (2002:74) for this prototypical model of sequencing rating behaviors.

al.'s "articulating scoring decision while summarizing and reinterpreting judgments" (2001:21).

Milanovic et al. (1996:95) proposed a similar sequence, but with a more detailed account of which textual aspects are assessed:

1. Internalize the marking scheme and interpret the tasks;
2. Scan for length, format, handwriting and organization;
3. Read composition quickly;
4. Rate by assessing relevance, development of topic coherence and organization, error distribution, command of syntactic complexity, and mechanics;
5. Reassess and revise if necessary; then
6. Decide final mark.

Although the above models of the sequence of decision-making behavior suggests a linear process, it is acknowledged in all the research cited that the process is often cyclical, characterized by recursions and interactions.

2.3.2.6 Variations in Decision-Making Behaviors

Despite the overall tendencies in decision-making behaviors outlined above, variations have been detected, testifying to Vaughan's (1991) contention that raters do not rate uniformly. Individual rating styles have been identified based on the importance raters attach to different assessment criteria and how they process textual information (Ecke, 2008; Milanovic et al. 1996; Sakyi, 2000; Sakyi, 2003; Smith, 2000; Vaughan, 1991). Variations have also been found to exist across rater groups related to such factors as their language backgrounds (e.g. Connor-Linton, 1995a; Kobayashi, 1992; Shi, 2001), academic and educational backgrounds (Cumming et al. 2001, 2002; Mendelsohn & Cumming 1987; O'Laughlin, 1994; Vann, Lorenz & Meyer, 1991) and experience (Cumming, 1990; Erdosy, 2004; Pula & Huot, 1993; Rinnert & Kobayashi, 2001; Song & Caruso, 1996).

For the purposes of rater training and rater selection, studies have also attempted to identify differences in behaviors between expert raters and non-expert raters. The results of these studies depend to some extent on how the notion of expertise is identified. Wolfe (1997) and Wolfe et al. (1998), for instance, defined an expert as one who can achieve high reliability levels while scoring written compositions. In these terms, expert raters go through a quick and efficient rating process, in which they focus on general features of the scripts and read large chunks of passages before judging them. Further, they have little personal engagement with the scripts, and they adhere closely to the scoring rubrics. Raters with lower expertise (i.e., raters with lower reliability scores), on the other hand, take a less fluent reading approach in which they often interrupt their reading processes to make preliminary judgments. Rather than making general references to textual features, less reliable raters make more specific references to these features, and they are less likely to adopt the values espoused in the scoring rubric.

Other studies attempting to detect decision-making behaviors associated with expert raters have looked at the raters' level of experience. Like Wolfe's highly reliable raters, experienced raters seem to be able to abstract their evaluations until the scripts have been interpreted (e.g., Cumming, 1990; Huot, 1993). What seems to distinguish experienced raters from highly reliable raters, however, is their tendency to make more personal comments and to rely on their internal set of criteria or "gut-level determination" (Pula & Huot: 1993:253) rather than the scale descriptors. So while experienced raters may rely more on their own background and experience than on the rubric, they have to set aside their personal engagement and values and adopt the rubric in order to achieve high reliability scores.

2.3.3 Summary of Raters' Decision-Making Behaviors in Independent Ratings

The growing body of research into the rating process has pointed toward certain trends in raters' decision-making behaviors in formal, independent writing assessments. The rating task appears to be complex and multi-faceted. Raters employ a variety of interpretation and judgment strategies while focusing on a range of textual features and monitoring their own rating behaviors. The process of rating student scripts involves creating an image of the scripts and

judging (the raters' images of) them. This process is not linear, but rather recursive in nature, involving a sequence wherein raters interpret student scripts while progressively building up their judgments of them. Raters appear to balance their attention across a wide range of textual features, although treatment of language errors seems to carry most weight in their decision-making for second-language writing.

When raters are presented with a rubric to guide their decision-making, the raters do not simply match the scripts to the scale descriptors in it. Rather, they seem to battle the demanding task of reconciling their complex assessments of the student scripts with the simple and abstract rating scales. The simplicity of rating scales fails to represent the complexity of the written scripts, so raters often feel frustrated, experiencing a tension between their roles as readers (and their natural inclination to draw on their own background to make meaning and judgments of scripts) and their roles as raters (requiring the pressure to adhere to the scoring rubric). So to reach a final score, raters often have to rely on their own expertise. Rather than representing a writing assessment construct then, scoring rubrics take on the reduced role of a justification tool, through which raters articulate their impressions.

Although prototypical decision-making behaviors may be determined, variations have been found across individual raters and across rater groups, and certain characteristic behaviors have been identified with expert raters and non-expert raters. Highly reliable raters and experienced raters exhibit a more fluent reading process than less reliable and novice raters in that they focus on general features of written scripts and are able to withhold their judgments until larger units of the scripts have been interpreted. Highly reliable raters and experienced raters, however, differ with respect to personal engagement and rubric adoption. Highly reliable raters refrain from getting too personally engaged with the writing and rely heavily on the rubrics for assigning a score; experienced raters engage personally with the scripts and rely more on their background and expertise than on the scoring rubrics.

2.3.4 Validity in Independent Ratings

The studies on raters' decision-making behaviors in writing assessment have made major contributions to the validation studies of writing assessment. They complement reliability-focused studies by shedding light on the construct of writing assessment as it is enacted by raters during the process of rating written scripts. It has been shown that raters do not assess students' written products in a uniform way: They vary individually, and they vary across groups. Common to all raters, however, is that they go through a complex process in which they attend to a variety of textual features and monitor their own rating behaviors while interpreting and building up judgments of written scripts. Studies of how raters consult a scoring rubric have shown that the rubric does not provide the raters with much substantial guidance in their decision-making process, and at times a rubric even leaves them frustrated in their attempts to align it with their intuitive impressions.

2.3.4.1 Validity of the Use of Scoring Rubrics

The basic issue in construct validity is whether an assessment measures what it is intended to (Davies, Brown, Elder, Hill, Lumley & McNamara, 2002:220) and thus "the extent to which we can make inferences about hypothesized abilities on the basis of test performance" (Bachman, 1997:256). So people must be able to infer from test scores information about the level of students' writing abilities. The above-mentioned studies indicate that a rubric-driven writing assessment procedure may not make it fully possible for stakeholders to interpret a test's results as an adequate indicator of students' writing ability. Scoring rubrics may not fully represent the writing assessment construct, and they may even, during the processes of rating, distort or be irrelevant to the actual assessments performed.

2.3.4.1.1 Failure of Rubrics to Represent the Writing Assessment Construct

If scoring rubrics rarely function as a direct contributor to raters' decision-making and merely serve to channel raters' "intuitive impressions of the text into a set of scores which they feel is publicly accountable" (Lumley, 2005:275), this has to "cast some doubt upon the idea that scales can assist us in understanding the constructs being measured by such ratings" (Lumley,

2002:266). As a consequence, established rating scales are somewhat limited in their validity. It is only raters and the rating community that will know what final scores mean. So stakeholders may be left poorly informed about a student's writing abilities based on assigned scores.

That the writing scale "remains lifeless and unrelated to language performance until used by a rater" (Lumley, 2005: 239) to justify or articulate a score leaves it in a very weak position. What occupies the central position in writing assessment, then, is not the scale, but the raters. Raters are the ones to focus on in a validation process because they are the ones who decide:

- which features of the scale to pay attention to;
- how to arbitrate between the inevitable conflicts in the scale wordings; and
- how to justify impressions of the script in terms of the institutional requirements represented by the scale and rater training (Lumley, 2002:267).

This conclusion confirms Pula and Huot's (1993) findings, which showed that "experience with scoring essays holistically, of having made the decision before, seemed to outweigh the influence and importance of any rubric per se" (1993:253).

2.3.4.1.2 Potentially Damaging Effects of the Rubrics

The above-mentioned studies have shown that raters tend to be conscientious about their rating jobs, trying hard to be loyal to a scoring rubric, but also that they have to rely on their own personal and professional backgrounds to interpret and evaluate students' scripts. This inevitably leads to a tension between their roles as raters and their roles as readers (Huot, 1990:55). On the one hand, raters experience pressure to conform to the predetermined scoring rubrics and to achieve high inter-rater reliability rates, and on the other hand they have to draw on their individual backgrounds and knowledge to make meaning of complex student scripts. In other words, in a writing assessment task, raters have to bracket their individual preconceptions and squeeze their complex and authentic reader responses into a predetermined set of abstract and context-free assessment criteria. This process may corrupt the authenticity of their reader

responses and as a consequence, their inferences about students' ability to communicate in writing may be compromised. As Cumming et al. put it,

> The simplicity of the holistic scoring method, and the rating scales that typically accompany it, obscures its principal virtue: reliance on the complex, richly informed judgments of skilled human raters to interpret the quality of students' writing performance (2002:68).

It must be noted here that some studies (e.g. Huot, 1993; Weigle, 2002) into raters' decision-making behaviors viewed training with rubrics more positively. Based on his study of novices and experienced raters' decision-making behaviors, which showed that experienced raters (i.e., raters who had previously been trained and worked with a scoring rubric) made more personal comments than did novice raters, Huot (1993) inferred that previous training with a rubric can free raters from spending energy on deciding how to make or channel their judgments. About a decade later, however, Huot became less enthusiastic about rubric-focused training procedures and harshly criticized traditional scoring procedures for their emphasis on "technology" and their underlying "positivist" epistemologies (2002:86).

2.3.4.2 Consequential Validity

Samuel Messick (1989), who has been very influential in establishing construct validity as the essential unifying component in validation, expanded the conception of validity to include the value implications and consequences of test use. He argued that validity encompasses not only the inferences that are made about the relationship between constructs and test scores, but also the value and ethical implications as well as the decisions and actions that are taken based upon those inferences and how they affect stakeholders (e.g., students, teachers, instructional programs, etc.). In other words, since Messick validity has come to subsume not just the accuracy and the appropriateness of the inferences that are made about the abilities one intends to measure, but also the decisions made based on those inferences and the appropriate use of those decisions as well as their impact and value implications. Taking these aspects into account, the empirical research conducted so far into raters' decision-making yields the following possible implications about the consequential validity of the traditional rubric-driven scoring procedure in independent writing assessments.

2.3.4.2.1 Ethical Issues

If stakeholders are not able to make confident inferences about students' writing skills from official scoring rubrics this may put the ethics of the testing procedures in jeopardy. Stakeholders themselves seldom have direct access to the actual assessment constructs that underlie the judgments of students' writing. Different societies, of course, have different concepts of what is ethically appropriate or correct, but both EALTA's (European Association for Language Testing and Assessment) and ILTA's (International Language Testing Association) code of ethics or good practice addresses the issue of stakeholders' access to relevant information about test scores[5].

2.3.4.2.2 Washback Effect

That raters are required to compromise (and sometimes even give up) their complex and subjective reader responses to a set of isolated and objective assessment criteria predetermined by a scoring rubric (e.g., to obtain high inter-rater reliability) may generate a negative washback effect on writing instruction. Kroll asks "what do our assessments teach?" (1998:223). A testing procedure that demands repression of personal reader responses to conform to a rubric may lead to writing instruction programs that ignore attention to the communication situation, including reader responses and personal engagement. Johnston, in particular, has warned against an overemphasis on objectivity in writing assessments, contending that, "The search for objectivity may not simply be futile. I believe it to be destructive" (1989:511). The argument here is that if educational institutions aim at moving students beyond the ability to regurgitate information, they must accept that "abilities like creativity, reflection, and critical thinking require a personal relationship with the subject" (1989:511).

[5] ILTA's Code of Ethics, Principle 1, annotation 6 write "Language testers shall endeavour to communicate the information they produce to all relevant stakeholders in as meaningful way as possible". EALTA's Guidelines for Good Practice p. 3, question 10 asks "Are the marking schemes/rating criteria described"

2.3.4.2.3 Value Implications and Epistemological Underpinnings

The values communicated by a rubric-driven, inter-rater reliability-focused assessment procedure appear at the expense of reader response. This approach values agreement and consensus over validity (particularly if the validity construct entails students' abilities to write for communicative purposes). Huot (1996, 2002) has even accused such assessment practices of representing an underlying reductionist and positivist notion of reality, primarily because they base their practices on objectivity through fixed, context-free language criteria, assuming "that there exists a reality out there, driven by immutable laws" (Guba, 1990:19) independent of the observer. Further, with the overarching goal of eliminating or masking differences (i.e., rater variance), such assessment procedures silence disparate voices and thus politically exude non-conformist values (Moss & Schutz, 2001).

These validity concerns essentially reflect the tension that raters' experience between their roles as readers and their roles as raters. They represent a serious construct validity dilemma: Should a scoring procedure aim at nurturing and acknowledging independent and often subjective human responses, or should it aim at making sure raters adhere to objective and prescribed reader responses, directed by scoring guides such as scoring rubrics and sample papers? Elbow's standpoint is clear: "Given the tension between validity and reliability – the trade-off between getting good pictures of what we are trying to test and good agreement among interpreters of those pictures – it makes most sense to put our chips on validity and allow reliability to suffer" (1991:xiii). This sentiment is echoed by Wiggins, who argued that "In performance testing of writing we are too often sacrificing validity for reliability, we sacrifice insight for efficiency, we sacrifice authenticity for ease of scoring" (1993:129).

2.3.4.3 A Clash of Two Theoretical Traditions

The validity dilemma, experienced by raters' tensions in actual writing assessment situations, is a concrete manifestation of conflicts between two major theoretical traditions: psychometric measurement, on the one hand, and composition theory, especially the notion of reader response, on the other. The main concern in psychometrics is to obtain measurable (usually quantitative)

evidence for validity, leading to an emphasis on score reliability. To obtain high reliability scores, objective scoring rubrics aim to control the disparate impact of personal experience and subjective interpretations that raters draw on to interpret and judge student texts. Composition theory, on the other hand, recognizes the importance of context and the individual, arguing that personal engagement and reliance on individual preconceptions (and therefore also reader variance) are necessary ingredients in reading and evaluation processes. The core of the argument is that reliability-focused assessment procedures "risk forcing potentially multidimensional rater responses into a single dimension" (Connor-Linton, 1995b:763). From this point of view, inter-rater reliability focused assessment procedures disregard or at worst corrupt the validity of scoring and run counter to the very construct that needs to be assessed. The two theoretical traditions, psychometrics and composition theory, may even contradict each other. This lack of reconciliation between the two fields is, in Huot's estimation, "a stalemate for writing assessment" (1996:552)[6].

It is, in fact, ironic that performance-based writing assessments were introduced to make writing tasks more authentic and allow independent and complex student responses, but the scoring procedures that typically accompany such tests work directly against their very purpose by attempting to control and reduce the complexity and the authenticity of raters' responses. This dichotomy begs the question of whether the practices of performance-based assessments should be judged apart from the theoretical principles that inform them.

2.4 Communal Ratings

2.4.1 Call for Alternative Writing Assessment Procedures

Because of the shortfalls of psychometric, rubric-driven assessment procedures, a number of scholars have called for alternative assessment procedures that prompt a (Re)Articulation of Writing Assessment (Huot, 2002: title) by putting raters into focus and accepting that "the age of

[6] The conflict between the "objectivist" paradigm of psychometrics and the "social constructivist" paradigm of literary and composition theory is reviewed in detail by Lynne (2004) and by Elliot (2005).

the rubric has passed" (Broad, 2003:4). In particular, calls have been made for assessment procedures that can integrate measurement theories and composition theories and that can accommodate an epistemological change in which validity or truth are not seen as absolute values but as context-dependent and based on the co-construction of meaning. The plea has been made by measurement scholars and composition scholars alike as a way to reconcile the two fields.

Pamela Moss, a prominent measurement scholar, has been particularly influential in calling for alternatives to purely psychometric assessment procedures. In her seminal article "Can there be validity without reliability" and accentuated in subsequent publications (e.g. Moss, 1996; Moss et al., 1998; Moss & Schutz, 2001) Moss (1994) has called for a hermeneutic approach to assessment. Recognizing that hermeneutics is not a unitary philosophy, Moss contends that philosophers within the hermeneutic tradition "share a holistic and integrative approach to interpretation of human phenomena that seeks to understand the whole in light of its parts, repeatedly testing interpretations against the available evidence until each of the parts can be accounted for in a coherent interpretation of the whole" (1994:7). What most significantly sets a hermeneutic approach apart from more psychometrically driven approaches to assessment is that it offers "a more pluralist approach and (that) allows dissensus to be represented and taken into account" (Moss & Schutz, 2001:37). Operationally, a hermeneutically-driven assessment practice would involve "interpretations of collected performances ...that privilege readers who are most knowledgeable about the context ...and that ground those interpretations...in a rational debate among the community of interpreters" (Moss, 1994:7). This form of assessment theory and practice thus sets itself apart from a purely psychometric approach to assessment by not only asking for complex language samples, but also by considering the particular context and by valuing and illuminating differences and disagreements among judges.

2.4.2 Communal Writing Assessment (CWA)

Communal Writing Assessment (CWA) has been viewed as a concrete manifestation of the current hermeneutic approaches to writing assessment (see especially Broad, 2000, 2003). Broad

and Boyd (2005) argued that CWA (along with portfolio assessment) epitomizes Huot's (1996, 2002) call for a new theory or paradigm of writing assessment that integrates measurement theory and composition theory.

In CWA procedures "two or more judges work to reach a joint decision on the basis of a writing performance" (Broad, 1997:134). As early as 1991, Belanoff stressed the importance of conversation among raters, "of pooling (of) individual judgments in the process of specific papers" (63), although she knew her suggestion was "fairly radical" (1991:64). This view differs markedly from that taken by psychometrically-driven writing assessment procedures that stress:

> Once live rating is under way, it is important to ensure that scoring is independent – that is, that raters do not see and therefore cannot be influenced by scores given by other raters...Whatever procedure is used, it is essential for the integrity of the scoring process that raters arrive at their scores independently, without reference to scores given by other raters (Weigle, 2002:131).

That raters can be influenced by each others' perspectives in CWA is seen as a major advantage from a hermeneutical point of view, not as a risk. Raters are put together precisely so that ratings can benefit from the possible synergy encouraged by more people working together. Ideally, everybody's voice is heard, and differences are accepted as a reality, and even as a benefit. While disagreements may be respected and accepted, the notions of reliability and standardization are not abandoned altogether in CWA. However, whereas differences in psychometrically driven assessment procedures tend to be reduced or eliminated, they are put to scrutiny and validated in CWA discussions. Standards are not ignored either. Raters are not left to apply their own idiosyncratic assessment criteria. They are expected to adhere to general standards. However, rather than having general standards predetermine the assessments, as is encouraged by the psychometric tradition, standards in hermeneutic CWA procedures are re-contextualized in the local assessments.

2.4.2.1 Scholarly Research into CWA

Research into raters' decision-making behaviors in communal writing assessment is in its infancy. Most scholarly work has stressed the theoretical advantages of such procedures, but studies grounded in empirical data have also begun to appear.

2.4.2.1.1 Theoretical Arguments for CWA as a Valid Assessment Practice

Because CWA accentuates raters' opportunity to validate their perceptions of student scripts against other human raters rather than (exclusively) against an abstract scoring rubric, such assessment practices, at least theoretically, offer promises of valid assessment. This type of assessment gives the raters an opportunity to obtain a nuanced assessment of the scripts in question. It also gives raters opportunities to develop professionally as raters and harmonizes with current composition and educational theories and their considerations of social contexts and individual agency.

2.4.2.1.1.1Sound Assessments

Resting on theories of social construction of meaning (Berger & Luckmann, 1966), a deeper understanding and appreciation of student scripts can be obtained through CWA because raters co-construct interpretations and judgments of students' scripts.

By bringing different perspectives into the construction of meaning through CWA, different sources can produce a rich and detailed understanding of written scripts. The interaction between professional raters can put their different perspectives to scrutiny as well as challenge, revise, and refine, and thereby validate, them. Some raters' perspectives may be better grounded than others: A solid line of argumentation supported by concrete evidence may be vouched for, whereas weak and illogical arguments with little or no evidence may be discarded. Thus, preconceptions and prejudices that may surface during the process of CWA are assessed and potentially judged as either disabling or inappropriate (too biased, too idiosyncratic, uninformed, dishonest, etc.) or productive and enriching to the assessment at hand. Mislevy emphasized the importance of uncovering such differences during assessments, stating that "when multiple

sources of evidence are available and they don't agree, we'd better have alternative lines of argumentation to establish the weight and relevance of the evidence of the inference being drawn ... We bear the burden of unraveling these possibilities" (1994:10-11). This contention has been echoed among writing assessment scholars in the past decade. For instance, Moss et al. claimed that such procedures as CWA "cannot only lead to an epistemologically sound, perhaps sounder, evaluation" (1998: 140), and Broad and Boyd insisted that deeper assessments are encouraged by diversity and not by "homogeneity" among assessors (2005: 16). Differences then serve an epistemic function and are not considered "measurement errors" as is the case in conventional psychometrics. A valid depiction of student writing is obtained precisely because raters approach student scripts in different ways using their different backgrounds and preconceptions to make meanings and perform evaluations. If raters disagree, it is not necessarily because they lack interest or knowledge, but because inevitably, human raters have different backgrounds, tastes, and standards.

During the process of unraveling differences, judges using CWA may come to realize that their initial perceptions need to be revised, and, as a result, a consensus is sometimes reached. It is not always the case, however, that the end result is a consensus. Positions may still differ as to what the most accurate interpretations and judgments are, and thus a compromise might be needed. To many hermeneutists (e.g. Gadamer, 1994; Hoy, 1994; Moss, 1996), lack of consensus is not a major concern, as "In real inquiries, agreement does not appear to be the essential telos of understanding, but a fortunate by-product" (Hoy, 1994: 265). An important goal in CWA is to gain an enlightened understanding of the differences and to move closer to what the raters view as accurate interpretations and judgments[7]. The process of putting different perspectives to scrutiny does not necessarily guarantee an accurate interpretation, but as Scriven has reminded educators, "this is the cross that all of us living in the new non-foundationalist age have to learn to bear!" (1972:31, cited in Moss, 1994).

[7] To emphasize their contention that the goal of an educational assessment discussion is not to assimilate different perspectives into consensus Moss and Schutz (2001:59) preferred the term "agreement" to the term "consensus" when referring to the end result of an assessment discussion. Agreement is defined as a case in which the participants "accept a particular conclusion in a particular context, what is agreed upon may actually be (and to some extent, always is) interpreted differently by each" (Moss & Schutz, 2001:59).

It follows that in order for CWA to be successful, considerations must be given to the choice of judges. Not only must they be knowledgeable about the language testing context, but they must also possess the ability to participate critically and coequally in rational debates and not let the more assertive voice dominate (Moss et al., 1998, Moss & Schutz, 2001). Agreement (whether it be consensus or compromise) should be reached on the grounds of rational persuasion, rather than oppressive coercion. However, echoing Habermas (1996), Moss and Schutz (2001:42) conceded that conversational power relations (influenced by culture, race, gender, and social class) affect any dialogue in subtle ways, and thus a dialogue completely free of any coercion would have to be an ideal that is rarely reached, but that nonetheless must be strived for.

It is not surprising that the scholarly work on such assessment procedures, and in many countries the actual implementation of such procedures, came in the wake of the increased focus on and uses of portfolio assessment. Just as portfolios (i.e., multiple language samples) are believed to enhance the content validity of writing assessments by illuminating different aspects of students' writing abilities (e.g., their ability to write in different genres, with different purposes, address different audiences and topics, etc.), communal assessment practices could be considered to enhance the scoring validity because multiple raters bring multiple perspectives to the interpretation and judgment of language scripts. In short, in the vein of improving the validity of the test task by bringing in multiple language samples to obtain a comprehensive and holistic picture of writing abilities, scoring validity is improved by bringing in multiple readers.

2.4.2.1.1.2 Contextualization

The opportunity given to raters to validate their perceptions of student scripts against other human raters and not just against a scoring rubric raises another aspect to the validity of CWA: the possibility of considering the writing context.

As Lumley observed, a scoring rubric for writing assessment is inevitably abstract because "it is written by others, using different eyes; it is public, formal, simplified, abbreviated and relativistic

in its language; it aims to be generalisable; and it cannot represent the individual perspectives of each rater, nor describe adequately each text" (2005:240). Nevertheless, raters in psychometrically driven assessment procedures are often required to adhere exclusively to scoring rubrics, and to do so they may be forced to ignore the actual writing context (including their own reader response) of the specific scripts. This requirement on raters is a manifestation of what Belanoff considered to be "the strongest myth of all, that it's possible to have an absolute standard and apply it uniformly" (1991: 55). As Belanoff further argued, this assumption implies "that there is some Platonic image out there of "good writing" and there is as a result a Platonic standard of writing which we can all learn to apply uniformly" (1991: 60).

In CWA rubrics are often still used, but raters themselves are the center of the scoring and validation process. The rubrics can be consulted, but they are expected to serve merely as a platform or a point of reference from which raters co-construct meanings and judgments. The rubrics are in such cases considered guidelines that raters bear in mind and respect but "reinterpret" (Moss, 1996:22) in the context of particular cases. In this way the official scoring criteria in the rubrics become flexible and subjective to inter-subjective interpretation and re-contextualization.

The process of arriving at a decision based on inter-subjectivity and re-contextualization may give the impression that raters are empowered to arrive at personal decisions about language ability. Moss (1996) and Moss and Shutz (2001) have not endorsed such relativism (or as some might argue, autonomy). Rather, they have advocated what they call a "dialectic between the contextualized understanding of local meanings and the distancing analysis of regularities" (Moss, 1996:22). Raters are not left to apply their own idiosyncratic assessment criteria. They are expected to adhere to general standards, but the standards are not expected to fully predetermine the raters' interpretations and judgments. Rather, they "co-determine" (Moss, 1996:22; Moss & Schutz, 2001:61) the raters' perceptions. Broad (2000) referred to this type of dialectic as "hermeneutic standardization" between the interpretive acts of raters and of regularized standards (e.g., the scoring rubrics).

If, as Huot declared, it is a "truism in current ideas about literacy that context is a critical component in the ability of people to transact meaning with written language (2002:101), the re-contextualization potentials inherent in CWA bring it closer to theoretical constructs of writing ability. Indeed, CWA reflects many theoretical and practical implementations of literary and writing programs (Huot, 2002). In many such traditions a deeper understanding and appreciation of writing are sought obtained by bringing in and scrutinizing different views of the texts. Not only are different perspectives "inevitable because they are part of a natural process of reading" (Stock & Robinson, 1987:105), but they are, in fact, crucial to a deeper understanding and appreciation of the texts. The notion of argumentation, too, is a strong element in the co-construction of meaning, appealing particularly to literary traditions and to expectations in writing programs as well.

In bringing practices of writing assessment closer to the writing construct, CWA to some extent, at least in principle, answers Huot's call for "a new set of theoretical assumptions and practices for writing assessment [which will] reconcile theoretical issues in measurement like validity and reliability with theoretical concerns in composition like rhetorical context and variable textual interpretation" (2002:82). In other words, giving raters the opportunity to rely on social context and their individual knowledge in responding to writing in assessment practices might relieve them of some of the tensions they experience between their roles as readers and their roles as raters.

2.4.2.1.1.3 Rater Development and Institutional Development

The opportunity given to raters to validate their interpretations and judgments against each other in CWA potentially leads to raters' professional development and to institutional development.

Raters using CWA are given opportunities to continually develop as raters. When CWA raters test out their assessment strategies, they receive feedback, which will, if they are reflective and responsive, lead to refined future assessment strategies. By conducting assessments with other raters, raters go through a continuous process of disputation and argumentation, which can lead

to a "heuristic for learning to construct validity arguments that contain strong consideration for alternate view as well as an understanding of how to create arguments that are compelling to various audiences" (Huot, 2002: 56).

As new raters and perhaps new assessment values enter the rating community, new ideas and values are gradually proposed and integrated, bringing to the fore a potential synergy between novices and experienced raters. Veterans can mentor novices, who in return can influence the veterans with their new ideas. Condon and Hamp-Lyons (1994) deliberately aimed at drawing on this potential for synergy when setting up reading groups consisting of veteran raters and novice raters for portfolio assessment.

Huot and Schendel pointed to not only the rater development opportunities in CWA but also to its potentials for institutional development because "assessment becomes a site where reflective teachers can shape future assessments as they reflect upon those in the past" (2001: 50). This view reflects that of Belanoff, who argued that "all the while recognizing that any evaluation system needs to grow from the strengths and initiations of individual teachers; it cannot be imposed from above – the standards must come from within the group and be constantly open to alteration and transmutation" (1991:64).

Durst et al. pointed to the advantages of CWA for students (one of the primary stakeholders of writing assessments): "If teachers have much to gain from these conversations, so do students. Surely, it is to their advantage to study with a teacher who is more broadly informed as a result of participating in these discussions" (1994: 296). This view was echoed by Belanoff, who stated that "The more we talk about evaluation with our colleagues, the better we'll become at giving feedback to our students on their writing and the better we'll be able to guide our students into making their own evaluations of all sorts of texts, including their own" (1991:64).

2.4.2.1.1.4 Reliability

The opportunity given to CWA raters to exchange their assessment strategies may potentially lead to increased inter-rater reliability. By gaining insight into each other's assessment strategies, CWA raters can reach a mutual understanding of each other's perspectives as well as of each others' assessment strategies and even of the official assessment criteria. This mutual understanding and appreciation of each others' strategies and of official assessment criteria can lead to acceptance and alignment of strategies. This potential is akin to Hare's (1976) claim that group validation of member judgments enhances cohesiveness.

As Moss noted, in less standardized assessment practices like CWA, the distinctions between reliability and validity blur because there will be "consonance among several lines of evidence supporting the intended interpretation" (1994: 7). Along the same lines Pula and Hula stressed that a community "permits raters to work as a group, achieving rating consensus, but at the same time retaining the individual and personal nature of their reading, which is so important to any description of the fluent reading process" 1993:260)[8].

In fact, the exercise of creating mutual understanding in communal rating sessions is not fundamentally different from what is commonly practiced in psychometrically-driven rater training sessions designed to enhance inter-rater reliability. Weigle (1994) showed that traditional norming sessions were, in fact, successful in getting the raters to rate more consistently, mostly by clarification of criteria, and modification of rater expectations.

Allen (1995:84) stated that the interactions in CWA sessions could, in fact, be considered 'norming' or 'standardizing', but he preferred the term 'shared evaluation', and Broad (2000:252) likewise used the term 'articulation', because the road to reliability in CWA does differ from the one in the psychometrically-driven norming sessions, which prepare for live rating sessions. In CWA reliability is built "from the ground up" where reliability is a result of "the richness as different perspectives are brought [together]" (Allen, 1995:83). In traditional norming sessions reliability comes about top-down and is a function of training for score agreement because standards are imposed from above (i.e., by official criteria or benchmark

[8] Pula and Huot (1993) did not investigate communal assessments, but they did emphasize the importance of community in training holistic scorers.

samples). Belanoff (1991) even likened norming or training sessions to "a form of brainwashing" (p. 59). She argued that raters are often treated like puppets who are not supposed to even question the validity of the standards imposed on them from above nor how to invoke these standards.

Rater discussions in CWA, on the other hand, form an integral part of the rating job, with the added by-product of potential rater development[9]. The purpose of such discussions is not for the raters to *discover* textual value by matching the scripts at hand to the scoring rubric. Rather, the job is for CWA raters to refer to official criteria to *construct* textual values on the basis of a collaborative interpretation of how well a student has accomplished a particular writing task (Broad, 1997). In Delandshere and Petrosky's words, "consistence, we argue, could be thought of as a process of confirmation rather than one of independent replication" (1994:16).

2.4.2.1.2 Summary of the Validity Potentials of CWA

The above review of the opportunities inherent in CWA to provide collaborative and contextualized assessments suggests that this type of writing assessment has the potential to enhance the validity of writing assessments, not just with respect to the inferences made about students' writing abilities based on their test performances, but also about the broader consequences of such assessment procedures.

2.4.2.1.2.1 Construct Validity

The essential concern in validity is whether inferences can legitimately be made about a student's writing abilities based on the students' test scores. There needs to be a close relationship between the construct of writing ability and the test scores. Messick argued that documentation of this relationship needs to be based on not only "empirical evidence", but also on "theoretical rationales" (1989:13). Although perhaps lacking in empirical documentation (see

[9] I deliberately choose the term 'development' when referring to CWA as opposed to the term 'training' used in psychometrically driven assessment to avoid the underlying behaviorist connotations of the word 'training'. The Concise Oxford Dictionary defines training as a process to "bring (person, animal, etc.) or come to a desired state or standard of efficiency, etc. by instruction and practice".

below) of this relationship, CWA can be said to be a step towards accommodating the theoretical construct of writing as a contextualized act related to current concepts of writing involving context-dependence as well as diversity and complexity of reader responses. Further, the notion that assessment criteria in CWA are not forced "from above" but codetermined by professional raters (including teachers) opens up new ideas to be generated and thereby adds a dynamic evolution to the assessment construct.

Reliability, too, is unlikely to suffer in CWA because CWA can bridge the gap between the psychometric need for consistency and the diversity and complexity valued in composition theory by pooling individual perspectives and prompting mutual clarification of assessment criteria to lead to rater alignment.

2.4.2.1.2.2 Consequential Validity

Compared to the psychometrically driven rubrics-focused assessment procedures, CWA exhibits promising potential with respect to consequential validity. This potential relates to the concrete impact of such assessment procedures as well as to their ethical, epistemological, and value implications.

2.4.2.1.2.2.1 Impact

CWA offers positive opportunities to many levels of education. Not only do the rater interactions in CWA provide raters with the possibility to continuously develop and refine their assessment strategies, but they can also benefit students as well as the writing program developers. The students who receive instruction from CWA raters can draw on the raters' extensive exposure to and experience with other readers' values in writing. Also, a likely positive washback effect would be a strong consideration of context and reader response in classrooms. On a more general level, the dynamics of assessment made possible by CWA can impact on institutional development as well as on assessment in general, providing a heuristic for continuous development within teaching as well as testing.

2.4.2.1.2.2.2 Value Implications and Epistemological Underpinnings

The value and epistemological underpinnings of CWA also differ from the more reliability-focused, rubric-driven assessment practices by adopting democratic values and a postmodern epistemology. Broad (2000, 2003), in particular, has stressed the democratic values underlying CWA. He argued that by letting minority voices as well as majority voices be heard, CWA relieves educators from choosing between "atomistic autonomy and oppressive community" (2000:254). As raters in CWA validate their interpretations and judgments against each other, rather than exclusively against an absolute scale, this process implies that truth is relativistic and not absolute. Different perspectives are articulated, and as Williamson pointed out, "explicitness about the process of decision-making through testing is perhaps the only basis for validity in a postmodern, postpositivistic world" (1993: 13). Wiggins likewise stated that "all assessment is subjective; the task is to make the judgment defensible and credible" (1994: 136).

CWA has even been seen as an enactment of Bohr's principle of complementarity (as developed in the fields of quantum physics) because not only does this theory "recognize the role of subjectivity in the collection and interpretation of data, it also abandons an obsession with reliability by acknowledging that differing experimental arrangements will sometimes yield contradictory evidence" (Broad & Boyd, 2005:11).

2.4.2.2. Empirical Research into CWA

Recent developments in scholarly studies of writing assessment offer substantial theoretical groundwork for a communal approach to writing assessment. These developments, of course, need to be backed up by empirical evidence (Messick, 1989:13). Empirical research into CWA practices, although still scarce and eclectic, has appeared in the last decade. Because CWA practices are considered alternative or innovative[10], much of the empirical research into such procedures has involved raters new to this type of assessment practice. Also, such studies have

[10] However, as will be shown below, educators in Denmark have practiced CWA in EFL wring tests for decades.

mostly been undertaken with EMT essays; research on ESL essays has involved just one study of which I am aware (Mohan & Low, 1995).

Being in their infancy, studies into CWA are exploratory and have not yet presented a detailed picture of raters' distinct decision-making behaviors, as has been the case with studies into rater behaviors in traditional rating sessions, where raters rate scripts independently. However, the few empirical studies that have been conducted in CWA provide an idea of the dynamics in such assessment practices.

2.4.2.2.1 Rater Dynamics in Communal Assessment

Pamela Moss, a strong advocate of hermeneutic educational measurement (described above), put her theories into practice with colleagues (Moss et al., 1998) to develop and evaluate a methodology for teacher licensure. Their empirical research did not investigate writing assessment (but rather focused on portfolio evaluation for teacher licensure in math), but they managed to document some of the qualities of negotiation when two or more raters collaborate to reach a judgment. Taking a qualitative approach to a first-time communal assessment session, Moss et al. (1998) framed certain issues related to the sequence and soundness of communal evaluations as well as the potential for professional development in such discussions.

2.4.2.2.1.1 Professional Development Potentials

Aiming to get at the developmental potential of communal assessment Moss et al. (1998) asked raters retrospectively about their communal rating experiences. The raters reported that they had had valuable learning experiences, which they would be able to use in their local school communities. In particular, they perceived potential for better teaching practices and empowerment to discuss assessment with other colleagues.

2.4.2.2.1.2 Sequence and Soundness of Evaluations (Critical and Coequal Participation)

Through analyses of tape recordings of one rater pair's interactions, Moss et al. (1998) attempted to identify raters' sequential rating process. They were particularly interested in the extent to which the raters' interpretations were being regularly challenged and elaborated upon as well as whether the raters participated coequally in the rating discussions.

The raters in their case study followed the basic order of the assessment criteria in the evaluation form presented to them. Their discussions fell into three separate phases. First, they exchanged information about what each had written down from their individual evaluations. When one person mentioned aspects of the candidate's performance that the other rater had not noticed, the raters would sometimes revisit their notes or extracts from the candidate's performance. However, due to time constraints, the raters tended to rely on their own notes instead of revisiting the actual performances of the candidates. In the second phase, the rater pair searched for and discussed counterevidence that could challenge their initial interpretations. In the third and final phase of their interactions the rater pair finalized their judgments. The final judgments were not a score but a summary statement intended to reflect their joint assessment. In their effort to arrive at a performance level decision, the raters did not revisit their notes or the candidates' performance. This was interpreted by Moss et al. (1998) as an indicator that their final decisions were based on "selective recollection" (p. 155).

Although Moss et al. (1998) did identify instances of seeking counterexamples to challenge initial assumptions, they noted a paucity of such instances and found that the raters "instead of confronting their assumptions...sometimes seemed to interpret evidence in ways that supported the assumptions they happened to already hold" (p. 152). On the basis of this observation Moss et al. (1998) concluded that there were limits to the dialectical movement between forming and challenging hypotheses with concrete evidence.

With respect to coequal participation in the rater discussions, few of the raters in Moss et al.'s (1998) study reported any significant inequalities in reaching a judgment. However, through observations and analyses of scripts, Moss et al. (1998) themselves noted some asymmetry in the rater discussions both with respect to writing roles (one rater usually took on the role of actually

writing down the final judgment summary) and speaking roles (although Moss et al. did not indicate how this asymmetry in speaking was manifested).

Despite the potential limits to the communal assessment practice seen in their study, Moss et al. (1998) were confident that the problematic issues could be addressed through rater training (or development) and that communal assessment practices could lead to sound evaluations and potential professional development.

2.4.2.2.2 Rater Discussions as a Score Resolution Method

Johnson et al. (2005) looked into communal assessment in writing tests, but only as a score resolution method and from a purely quantitative approach. The purpose of their study was to find out whether discussions formed a sound alternative to averaging scores in score resolution. They claimed to be investigating the "accuracy of scores" (2005: 117) and thereby the "validity" (2005: 126) of such score resolution cases by focusing on (a) how closely the discussion-based scores and the averaged scores approximated the scores produced by a validation committee and (b) the extent to which raters engaged equally in the discussion process or whether there were signs of dominance or deference.

2.4.2.2.2.1 Accuracy of Scores

In their attempt to get at an "accurate" score by calculating the extent to which the discussion-based resolution scores agree with the scores set by the validation committee, Johnson et al. (2005) took a quantitative and psychometric approach to validity (emphasized by the experts in their validation committee being selected for their inter-rater reliability). Although they argued that "reliability requires parallelism or interchangeability of assessment conditions...and that "raters [in their study] are not interchangeable with members of the validation committee" (p. 124), this approach to investigating accuracy of scores is very much like investigating inter-rater agreement. Regardless of whether one characterizes this approach as a focus on inter-rater reliability or not, the researchers took a quantitative approach to validity by focusing on

agreement level and adopting a viewpoint that assumes an absolute truth (i.e., the scores from the validation committee). This positivistic approach runs counter to the very purpose of a discussion-based, CWA approach, whose claim to validity relies on the synergy of two or more raters co-constructing meaning in a particular context.

2.4.2.2.2.2 Dominance

In their focus on the risk of dominance and coercion and their effect on validity, Johnson et al. drew on Moss' stated requirement of CWA raters that they "Remain equally…engaged in the dialogue rather than acquiescing to the more assertive voice or the more comfortable decision" (1996:26). However, rather than focusing on conversational dominance (as Moss herself intended and investigated in Moss et al, 1998 – see above), Johnson et al. (2005) focused on score dominance. They assumed that if raters are equally engaged in the rating discussions, "it also seems reasonable to expect that the discussion scores would agree equally with the scores from the original raters" (p. 126). Relating this to validity, they claimed,

> If a majority of discussion scores agree with the original scores of one of the raters, it offers initial evidence that the view of that rater might be dominant in the discussion process….Hence, one could argue that there appears to be the possibility that a dominance effect in the use of discussion could result in scores of reduced accuracy, and thence, reduced validity (p. 126).

As Johnson et al. (2005) themselves acknowledged, validity would not be reduced in cases where the rater of lesser expertise defers to the other rater in recognition of that other rater's expertise. However, while recognizing the possibility of dominance having a positive effect on the score outcome (and on rater development in general), Johnson et al. (2005) nevertheless chose to investigate score accuracy on the basis of score dominance, thereby ignoring the potential benefit of raters being positively influenced by one another.

Whether or not one agrees with Johnson et al.'s (2005) psychometric approach to determining the validity of discussion-based scores, they must be credited for at least looking into the validity potential of this approach to score resolution. Based on their approach to validity (by making

correlations with expert scores and lack of score dominance), their results showed that the "accuracy" of the discussion scores and the averaged scores were "about the same" (p. 138) when scored holistically. The same was the case with the use of an analytic rubric for the domains of Style, Conventions, and Sentence Formation. However, the discussion based-approach fared slightly (i.e., showed no statistical significance) better than the averaged-based approach with respect to the domains of Content and Organization in their study. In no instances was the correlation between the averaged-based score and the "expert score" higher than the correlation between the discussion-based score and the validation committee.

With respect to dominance, Johnson et al. (2005) did find that discussion-based scores agreed more frequently with the original score of one of the raters, "indicating the possibility of rater dominance or deference" (p. 139). However, the statistics did not exceed the critical value, so there was "insufficient evidence to demonstrate that any dominance found in the sample can be attributed to anything more than chance" (p. 140). Further, it was reported that in instances of score dominance (when the discussion-based score agreed more frequently with the original score of one rater), there was little influence on "score accuracy" (p. 140), as in such instances there was a .86 correlation between the discussion-based score and the expert-criterion related score.

On the basis of these results, Johnson et al. (2005) concluded that a discussion-based approach should be used in high stakes assessments, but they called for further research to substantiate this claim. They also indicated that besides a sound evaluation of the essays in question, discussion-based scoring could be an important method of professional development.

2.4.2.2.3 Professional Development Potentials of CWA

Other studies that have investigated CWA have emphasized the professional development potential inherent in such assessment practices. Because these studies have used newly established CWA practices as a frame for their studies, conclusions cannot be drawn about the

long-term effect of such practices. However, certain potentials for professional development have appeared.

Allen (1995) ventured to invite teachers of different programs using portfolio assessments to assess each other's portfolios in light of the individual programs. The "shared assessments" took place in an internet-based discussion forum and through email exchanges. Despite initial anxieties about criticizing each others' programs, the teachers/raters showed "an ethic of disciplined collaborative inquiry that encourages challenges and revisions to initial interpretations" (p. 68) and a surprisingly high level of agreement among the raters (82.5%). Allen interpreted these collaborative exchanges as a constructivist scheme paving the ground for a hermeneutic circle with major potentials for professional development. He further hypothesized that a high level of inter-rater agreement would be reached if teachers were left to discuss standards amongst themselves. Allen (1995) argued that the rater development taking place in these CWA sessions were essentially different from what goes on in psychometric training sessions in that in the rater discussions in his study did not involve the "same drive to "calibrate" our readings so that we all read the same way" (p. 81). Apart from the potentials for professional development, Allen also believed rater discussions "can lead a local assessment procedure to increased fairness" (p. 84) although he doubted that such an assessment procedure could substitute for large-scale assessment programs.

Durst et al. (1994) reported on the potential gains in professional development when raters engage in discussions to form a judgment on portfolio writing. They had "good, solid teachers" (p. 295) engage in CWA of portfolios. Although initially the aim of the rater discussions in their portfolio writing assessments was to determine whether students should pass or fail particular portfolios - and the raters, the good, solid teachers, did not always agree - "what emerged was an opportunity for teachers to reflect on ways in which their own standards can evolve and be modified in this process of a portfolio conversation" (p. 296). So even though the raters' readings of the portfolio were not always identical, they were indeed "mutually illuminating" (p. 292). Durst et al. (1994) also noticed that the potentials for professional development inherent in such communal writing assessment conversations did not just concern a refinement of the raters'

assessment strategies, but also that the raters had the added opportunity of serving as resources for each other with respect to their teaching situations.

Condon and Hamp-Lyons (1994) had similar experiences with a different group of raters. They set up communal rating sessions with the purpose of providing raters with faculty development. They found that their raters, after having participated in such reading discussions, "shared their theories and practices to a greater extent than before" (p. 284). Novices and veterans benefitted from being grouped together in that the novices learned from the experience of the veterans, who in turn felt inspired from the novices' new energy and ideas. Condon and Hamp-Lyons' overall impression of the establishment of such discussion groups was that they "not only improved the quality of the assessment, but had a positive impact on the teaching/learning environment in the course as well" (283). This led the researchers to conclude that such "assessments were more accurate and more fair "(284).

Although not focused on standardized test as such, but rather writing pedagogy in general, Nixon and McClay (2007) conducted a case study of the interactions of three elementary school teachers and concluded that these dialogues "promoted thoughtful grading practices and encouraged instances of objective reframing of beliefs and assumptions about writing pedagogy" (p. 149).

2.4.2.2.4 CWA and Standard Setting

Other studies that have looked into the rater discussions in CWA have focused on the aspect of standard setting.

Mohan and Low (1995) asked teachers of the same courses to collaboratively define assessment criteria and subsequently apply those constructed criteria consistently to ESL essays. The researchers detected a large amount of disagreement among the raters, who were all new to CWA, particularly in relation to different beliefs about the separation of language and content. This was most pronounced once the actual scoring began, producing a tension among the

teachers, which "caused the teachers to question the validity and value of this approach" (pp. 28-29).

Nevertheless, lengthy rater discussions in which the raters were "reflecting on each other's perspectives, clarifying their ideas, going back to the composition to reach agreement and understanding" (p. 30), as well as raters' retrospective remarks such as "marking together helps me understand the criteria more clearly" (p. 29), led Mohan and Low to infer that CWA "helped them [the raters] to broaden their perspective and, at the same time, to recognize the need to learn more about that relationship [between language and content] as they continue to teach and evaluate language and content tasks" (p. 30). Mohan and Low concluded that such assessment practices pave the way for developing shared meaning over time and envisage that they "can help to produce fairer and more consistent evaluation" (p. 31).

2.4.2.2.5 Broad's Arguments and Research

Bob Broad, a passionate adherent of CWA (and a fervent critic of psychometrically rubric-driven assessment procedures) has published widely on the benefits of communal assessment practices in portfolio writing assessment for placement purposes. In 1997 he documented the "rhetorical and political dynamics" of how raters with different areas of expertise (instructors, teachers, and administrators) interact in communal assessments and found a mixture of democratic and autocratic processes.

In 2000 Broad went on to explore those dynamics in a standardization process grounded in CWA. In their aim to reach standardization, the raters in Broad's (2000) study struggled and "experienced multiple breakdowns" (p. 213), which were manifested most clearly in two areas: textual representation and evaluative subjectivity.

2.4.2.2.5.1 Textual Representation

The raters found it difficult to find representative texts. In fact, due to a high degree of rhetorical and contextual differences among the texts the raters even found it difficult to distinguish between representative and unrepresentative texts.

2.4.2.2.5.2 Evaluative Subjectivity

With respect to aligning their perceptions of what good writing is, the raters, although eager and striving hard to standardize themselves, found it frustrating to align their subjective perceptions to a common standard. It did not make the standardization process easier when the raters' standards shifted from one script to another and from one context to another. In their attempts to align their standards, the raters made efforts to treat different and sometimes extreme perspectives as legitimate and were even at times persuaded to change their perspectives, although they were sometimes uneasy with the changes. While the raters in Broad's study generally interpreted these crises as failure, Broad chose to not evaluate these scenarios from a psychometric perspective, but rather from a hermeneutic perspective. He interpreted the raters' tireless efforts to maintain their diversity and complexity while at the same time striving for standardization and coherence as a positive trait of their assessment community.

2.4.2.2.5.3 Dynamic Criteria Mapping

Broad (2003) emphasized the significance of publicizing the voices of the different evaluative perspectives; the voices of the majority as well as the voices of the minority. Mapping and subsequently publicizing what assessment criteria raters invoked to defend and negotiate their judgments in discussions with other raters would, Broad argued, form an honest picture of "What we really value" (2003). The criteria put forth in rater negotiations would indicate what is important to them. This, he argued, "mov[es] us beyond rubrics, traditionally the main obstacle to telling the full and true story of how writing is valued" (2003:122).

To unveil the assessment criteria invoked in a CWA program, Broad (2003) developed and applied his Dynamic Criteria Mapping (DCM) model with an existing writing program. He

recorded the assessment criteria that appeared in the raters' debates and reflections during their communal assessments of students' writing.

Broad (2003) divided the assessment criteria up into Textual Criteria and Contextual Criteria. As the names imply, Textual Criteria involve criteria directly related to student scripts, whereas the Contextual Criteria concern criteria not directly related to the scripts, but rather criteria pertaining to the assessment and the instructional context. The Contextual Criteria concern references to standards or expectations, fulfilling the assignment, learning or progress, plagiarism, essay vs. portfolio, compassion for writer, and the like. Of the Contextual Criteria, fulfilling the assignment seemed to be the most significant, often serving as a "gateway criterion" (p. 81).

Broad divided the Textual Criteria into two main subcategories: Textual Qualities and Textual Features. Textual Qualities refer to aspects of the reading experience and include aspects such as significance/development/heart; interesting/lively/creative; thinking/analysis/ideas; or unity/harmony/connection. Textual Features, on the other hand, relate to elements of the text and can be physically pointed to more easily. They include aspects such as mechanics/conventions/mistakes/errors; grammar; punctuation; spelling; content/topic; sophistication; variety; clarity; detail/description/ examples/dialogue; length/amount. Not surprisingly, as the venue for Broad's (2003) study (or criteria mapping) was a university EMT portfolio program, administrators and instructors discussed Textual Qualities somewhat more than they discussed Textual Features. However, despite the fact that the focus of the study was EMT writing scripts and that neither program documents nor comments from instructors and administrators indicated a particular emphasis on the features of Mechanics (error, editing, mistakes, conventions, length), Broad found a "clear dominance" (2003:62) of attention to such aspects in rater discussions. An explanation put forth for this disproportionate attention to Mechanics was that "Mechanics were safer to talk about than other, more complex, potentially more contentious aspects of rhetoric" (p. 63).

Length, in particular, was discussed to a great extent and was often used as the "bottom line" criterion for the final judgment. However, "as if, knowing how powerful and yet superficial a criterion Length could be, they [the raters] wished to prevent it from shading out other, more significant rhetorical values" (p. 68) and so consistently combined it with other Textual Features, such as Learning Progress.

Broad's DCM was intended for application to a local portfolio writing program, so the results of What [They] Really Value (2003, title) cannot be fully extended to other programs[11] (especially his context-bound criteria), but his in-depth analyses of the various assessment criteria do provide general information about what was valued in the one context of writing.

2.4.2.2.6 Summary of Empirical Research on CWA

The research conducted on CWA is sparse and eclectic. It has, however, uncovered some of the potentials and intricacies of the interactions taking place in such assessment practices. These point toward CWA as a promising method to enhance scoring validity in two respects: to improve validity of scores and to produce positive impact on rater development. The studies conducted are all recent and only report on innovative rather than established practices.

2.4.2.2.6.1 Rater Development Potentials

Most of the studies into CWA, or communal assessment in general, point towards great potential for rater development. Through collaborative inquiries raters seem to take the opportunity to challenge and revise their initial interpretations and judgments, and they report understanding assessment criteria more clearly (than they would with independent, conventional assessment methods). Further, inter-rater agreement seems to be relatively high in such practices. The developmental potentials pertain not only to assessment but also to teaching contexts as the raters participating in the studies reported that they used each other as resources for subsequent

[11] That Broad's DCM is not fully generalizable is not a flaw of his model. Rather, it "demonstrates how evanescent and particularized evaluation is" (Belanoff & Denny, 2006:135).

teaching. With these promising prospects for rater development, CWA can be said to have a potentially high degree of value for consequential validity.

2.4.2.2.6.2 Sound and Valid Assessments

From a hermeneutic point of view a sound and valid assessment is reached if raters engage critically and coequally in rater discussions, exploring and challenging each others' perspectives. The research conducted so far has indeed pointed to critical and coequal rater engagement in CWA interactions. The raters have seemed to challenge their initial interpretations (although not actively seeking counter evidence), and they have revisited their notes and, to some extent, also actual student performances. However, perhaps due to time constraints, the raters were perceived to make limited use of concrete evidence from student performances when forming and challenging their own hypotheses. Having identified this as a problematic area, researchers such as Moss et al. (1998) were nevertheless confident that this could be dealt with in rater training (or development).

With respect to coequal participation, the raters in these studies themselves reported no or little inequality, although some asymmetry was noticed by the researchers (e.g., the fact that one rater usually wrote out final judgment notes). Again, the researchers considered this to be a minor problem that could easily be addressed in rater training (or development).

Approaching sound and valid assessments from a more psychometric perspective (e.g., by correlating scores to the scores of a validation committee and viewing coequal participation in terms of score dominance) CWA appeared to exhibit no significant dominance nor decrease in accuracy compared to procedures of averaging scores.

2.4.2.2.6.3 Contextualization

With respect to contextualization of standardized criteria and taking social or educational contexts into consideration, including the exchange of reader responses, research into CWA has showed that CWA raters do seem to consider such contexts.

The raters in these studies did attempt to contextualize the standard assessment criteria, when such a set of criteria was present. When there were no standardized criteria to draw on, the raters put considerable effort into standardizing their different perspectives. When considering other raters' viewpoints, the raters respected each others' viewpoints, and sometimes even changed their own assessment criteria. However, tensions occurred, and the raters often felt frustrated with the attempt to standardize or align their different perspectives, sometimes questioning the validity of a communal approach. The frustrations were also a result of the difficulty the raters experienced finding representative texts for standardization. Although the raters seemed discouraged by what they saw as a failure to find representative texts and to align their viewpoints, researchers such as Broad (2000) interpreted this tendency as a budding hermeneutic assessment practice in which it is recognized that raters have different perspectives and that textual representation is hard to establish because assessment can only be a contextualized act. Only one study (Broad, 2003) looked at the actual assessment criteria invoked in CWAs. This study showed that the raters attended to textual as well as contextual features, although textual features such as mechanics, especially length, seemed to have a major impact on the final assessment.

In sum, empirical research into CWA points towards sound, contextualized assessment with rater development potentials. However, research is still scanty, and further research needs to complement and expand on the studies conducted so far. More research into CWA will add to the validation of this type of assessment and provide useful input for professional development as proposed by Moss et al. (1998).

2.5 From Independent Ratings to Communal Ratings

Research into raters' decision-making has provided valuable insights into what textual and contextual features raters invoke in their assessment efforts and how they conduct themselves in their rating process when rating students' written scripts independently. When conscientiously attempting to apply a predetermined rubric, some studies have suggested that such assessment tools are unhelpful and sometimes even counterproductive in assisting raters in assessment tasks (Broad, 2000; Elbow, 1993; Lumley, 2002, 2005). Raters sometimes apply their own internal set of criteria either as a supplement to official scoring guides or from their own beliefs about the value of writing qualities (Sakyi, 2003; Vaughan, 1991).

CWA has been proposed as an alternative to rubrics-driven assessment practices in which raters must rate independently to avoid influencing one another. Although CWA practices offer promising potentials for validity and professional development, little empirical research has been conducted in such forms of assessments to document these potentials. The few studies that have been undertaken in CWA have confirmed the theoretical assumptions of rater development potentials and have hinted at potentials for sound and contextualized assessments. This small set of studies, however, is exploratory and has not yet been able to present a detailed portrait of raters' distinct decision-making behaviors in these contexts. No study so far has attempted to map raters' decision-making behaviors in CWA in its entirety, tracing their decision-making behaviors from independent ratings to communal ratings nor to document how assessments may evolve from one session to the other. In the present thesis study I have intended to do that.

2.5.1 The Present Thesis Study and Research Questions

To contribute to the establishment of a theoretical framework of CWA, I have employed empirical data to verify previous findings as well as explore in detail the decision-making behaviors of one set of CWA raters. The research has traced the raters' decisions when they formed their individual interpretations and judgments in their independent rating sessions to when they challenged these perspectives in discussions with a co-rater during communal rating sessions. My intent is to map and discuss the raters' distinct decision-making behaviors and chronicle the progression of these behaviors as the raters move from their independent rating

sessions to their communal rating sessions. The specific research questions that emerged from this research focus are:

1. What are the raters' distinct decision-making behaviors and what is the sequence of these behaviors in the independent rating sessions and in the communal rating sessions?
2. How do the raters distribute their attention to the official assessment criteria in the two rating sessions and how does this distribution of attention differ from one session to the other?
3. To what extent do the raters engage equally in the communal rating sessions?
4. What are agreement levels and score ranges in the two rating sessions and how do they compare to one another?
5. What are the raters' perceptions of CWA in general and in relation to the specific CWA they have just practiced?

Chapter 3
Methods

3.1 Purpose and Organization of the Chapter

With the purpose of mapping CWA raters' decision-making behaviors and addressing the specific research questions I conducted an empirical study with 20 highly experienced CWA raters. This chapter describes the background to this empirical research context and the procedure for collecting and analyzing the data. It first outlines the venue of the study, the HHX EFL written exam in Denmark, which has a long and well-established tradition of employing communal assessments. It then describes in detail how the data were collected, coded and analyzed. The data sources were built on verbal reports in the form of concurrent think-aloud protocols and recordings from rater discussions in independent ratings and communal ratings respectively. These sources were supported by retrospective reports and score distribution. The verbal reports were transcribed and put into quantifiable form by segmenting and coding the data. The final coding scheme made it possible to map the raters' decision-making behaviors in their independent rating sessions as well as in their communal rating sessions, and it facilitated a comparison of the raters' behaviors in the two sessions. More specifically, the coding scheme made it possible to map and compare the raters' interpretation and judgment strategies, their self-monitoring and interactional foci as well as the assessment criteria they invoked to interpret and judge the student scripts in the two rating sessions, including the extent to which they considered the official assessment criteria. The communal rating sessions were further examined for signs of equal engagement by focusing on score dominance and conversational dominance. The raters' retrospective perceptions of CWA as well as their score ranges and agreement levels were recorded to shed further light onto the raters' decision-making behaviors and the progression of their behaviors.

3.2 Background to the Research Context

The venue for my research was the CWA procedures practiced to assess the results on the written component of the EFL exam of the Danish HHX (Højere Handelsskole Eksamen = Higher Commercial Examination) from 2004. This is an official, national exit exam for students completing upper secondary school (high school) with focus on business and socio-economic disciplines in combination with foreign languages and other general subjects. A description of the policies and practical procedures surrounding this exam is provided below, along with a brief history of the origin of communal assessment in Denmark. Denmark has a long tradition of communal assessment procedures throughout the educational system in all types of subject areas. Using CWA in Denmark as the research context of my study then paves the ground for investigating CWA as a well-established assessment practice with raters highly experienced in CWA.

3.2.1 Communal Assessment – a Tradition in the Danish Educational System

Denmark has a long (and from an international perspective, a unique) tradition of communal assessment (Plischewski, 2003). It dates back to more than a hundred years ago (Danmarks Evalueringsinstitut, 2005a; Haue, 2000) and is an important element of quality assurance[12]. It is rooted in having external examiners guarantee that students receive equal and just treatments as well as ensuring quality and national standards (Danmarks Evalueringsinstitut, 2005a; Haue, 2000). It is still practiced throughout the educational system for national, standardized exit exams, from elementary school (9 or 10 years of schooling) over secondary schools (typically 3 additional years of schooling) to higher education[13].

The common objective of upper secondary schools (high schools) such as HHX, STX, HF, and HTX[14] is to prepare people for higher education, and as most institutions of higher education in

[12] For more on quality assurance in the Danish educational system, see eng.uvm.dk/factsheets/quality.htm?menuid=2505

[13] Private and public schools are subject to the same exams and exam procedures.

[14] HHX (Højere Handelseksamen) focuses on business and socio-economic disciplines in combination with foreign languages and other general subjects; HTX (Højere Teknisk eksamen) focuses on technical and scientific subjects in combination with general subjects; STX (Studentereksamen) and HF (Højere Forberedelseseksamen) consists of a broad range of subjects in the fields of humanities, natural science and social science. Although the objective of the

Denmark do not have entry exams, the exit exams from these schools can be regarded as indirect entry exams to higher education. The average score of the exit exams and the students' annual grades, which are based on the students' performance during the school year, is calculated and determines whether the upper secondary school diploma (high school diploma) can be earned and whether the students can be admitted to a particular college or university. The upper secondary school exams are thus considered high-stakes exams.

The Ministry of Education formulates all written examination tasks and appoints external examiners (raters) for both oral and written examinations. For the oral exams the classroom teacher and an external examiner judge the students' performances in communal ratings. For the written exams the students' performances are exclusively assessed by two external examiners in a communal rating[15].

3.2.2 The HHX Written EFL Exam Context

The written EFL component of the HHX exam is an integrated exam that tests "the student's ability to treat a topic in a coherent way. The test must give the student an opportunity to independently structure a topic in producing a lengthy text"[16]. It is independent of the curriculum and can thus be considered a proficiency test. It consists of 3 parts: a summary of an English text (Part I); a translation from a Danish text into an English text (Part II); and a composition (e.g., an essay, a report, a speech) based on the same topic as the accompanying texts (Part III). Part I counts 25% towards the final score, Part II counts 25% towards the final score, and Part III counts 50% towards the final score. Students have 5 hours to complete the three parts. All the three exam questions take a starting point in an accompanying reading packet (the two-page English newspaper article to be summarized – Part I of the exam and the two-page Danish

---

vocationally oriented secondary education is not to prepare for higher education, they are also subject to external examination based on communal rating (http://eng.uvm.dk/publications/factsheets/htx08o_000.htm?menuid=2515).

[15] See http;//eng.uvm.dk/factsheets/?menuid=25 for more information in English on the Danish educational system and examinations.

[16] Translated from www.uvm.dk ("Formålet med opgaven er at danne grundlag for at bedømme elevens evne til fyldigt at behandle et emne i en sprogligt sammenhængende form. Opgaven skal give eleven lejlighed til selvstændigt at strukturere og formulere sig i en længere tekst") (www.us.uvm.dk/gymnasie/erhverv/eksamen/Vejledninger/vejl99en.html).

newspaper article to be translated – Part II of the exam). The exam questions and the accompanying reading packet are never reused, so for each exam a new set of materials is constructed. (See Appendix A for the entire exam packet used for the exam studied in this thesis.) The test developers that construct these tests are current teachers of HHX.

3.2.3 Students

At this exam, the test-takers are usually 18 to 20 years old and have had 8 to 9 years of instruction previously in English as a foreign language (335 hours during their HHX[17] years and approximately 570 hours during their elementary school years[18]).

3.2.4 Raters

As with most raters of official exams in Denmark, the raters for the HHX written EFL exam are employed on contract by the Danish Ministry of Education. These raters are experienced teachers teaching at the Handelsskole, the upper secondary school leading up to the HHX[19]. They come from all over Denmark and are nominated by their local schools and selected by The Ministry of Education to become members of the national rater corps. The national rater corps for the HHX's written EFL exam consists of approximately 80 raters, about 10 per cent of whom are replaced each year (Karsten Sielemann[20], 2004-personal communication; further information, see also Danmarks Evalueringsinstitut, 2005b for report on external examination procedures).

[17] www.ug.dk/vejledningsportal/Elementer/Guide%20til/Artikler.aspx?article_id=artik-hhxengelska
[18] http://us.uvm.dk/grundskole/generelinformation/vejlendendetimetal/timetal/pdf
[19] The Ministry of Education can choose to also include in the rater corps raters that are not currently teaching at the Handelsskole (e.g. other relevant stakeholders such as representatives from higher education). The year that the data for this study were collected the rater corps consisted of Handelsskole teachers exclusively.
[20] Karsten Sielemann is the chair of the national rater corps for HHX – English.

3.2.5 Training

Except for the general orientation at the Annual CWA Gathering described below the raters do not undergo specific training or a calibration process to be part of the rater corps. They enter the national rater corps based on their previous teaching experience and recommendation from their local schools. Their experience in the national rater corps with its CWA is considered continuous training (or development, rather). The procedure of replacing approximately10 per cent of the raters each year ensures monitoring of new members and exposure of new ideas from new members to senior members.

3.2.6 The Annual CWA Procedure

Like the other standardized exit exams for the other upper secondary schools (HTX, STX, HF), the exams for HHX are held once a year in the late spring. All students around the country who sit for these exams do it on the exact same day at their respective schools. Shortly after the students have completed their exams, copies of their exam scripts are sent to the chair[21] of the rater corps, who distributes the scripts among the raters of the rater corps. Each student script is rated by two different raters. Each rater receives approximately 100 student scripts (corresponding to student scripts from 3 to 4 different schools). No rater will be assigned to rate any students from his/her own school.

The actual rating takes place in two different sets of sessions: The independent rating sessions and the communal rating sessions.

3.2.6.1 Independent Rating Sessions

After having received the student scripts the raters are given approximately two weeks to rate the scripts independently in preparation for the final scoring in the communal rating sessions. They

[21] The chair of the rater corps is an experienced teacher and rater of the HHX. He is appointed and employed by the Ministry of Education.

are free to rate the scripts at any time or place and can spend as much time as they like on each script[22].

3.2.6.2 Communal Rating Sessions at the Annual CWA Gathering

After the independent rating sessions all raters of the national rater corps meet for their communal rating sessions at a communal rating gathering. This occurs 2 to 3 weeks after the students have sat for their exams and takes place at one location nationally, usually in Funen, which is centrally located in Denmark.

As an introduction to the actual communal rating sessions, the chair of the national rater corps hosts a brief preliminary discussion about possible challenges related to the current exam questions and student scripts. Directly following this debriefing the raters are divided into communal rating dyads: the two raters that have been assigned to rate the same set of student scripts in the independent rating sessions now sit face-to-face to co-rate in the attempt to reach a final score on each of their student scripts. The dyads are not restricted to resolving score differences. The raters can, if they feel the need, discuss student scripts to which they have given the same score[23]. All raters from the national rater corps are in the same room, so in case of severe disagreement between two raters, they can ask for a third opinion[24]. Each rater dyad rates about 30 student scripts together, which means that they rate with approximately three different raters at each communal rating gathering. Although each rater of the national rater corps has taught students sitting for this exam, they are never assigned to rate students they themselves have taught. Thus they can be said to be well-informed, but distanced, raters. To increase the circulation among the raters, each rater is assigned new co-raters every year.

[22] They are paid per script.
[23] This is unlike the practice studied by Johnson et al. (2005), in which discussions were restricted to score resolution cases.
[24] This happens only in very rare cases (Karsten Sielemann, 2004 – personal communication).

3.2.7 Scoring Rubric

The raters are asked specifically to make a holistic, not an analytic, assessment of the student scripts. Assessment criteria, a scale, and accompanying student script samples are presented to guide the raters in their rating process. All these materials are available to the public on the website of The Ministry of Education (www.uvm.dk). The rubric was developed by a small group of teachers teaching for HHX and finalized by the chair of the national rating corps for HHX – English.

The scripts are rated on the well-established "13-scale". This scale was introduced in 1963 (Petersen, 2006; Undervisningsministeriet, 2004) and has until recently been used all through the Danish educational system in all subjects[25]. A new scale was introduced in 2006 and 2007, i.e. after the collection of the present data. The 13-scale is a 10-point scale, with 13 being the highest score, and 0 being the lowest score. There is no 12, 4, 2, or 1[26]. The general scale descriptors for this scale are presented in Appendix B1).

The 13-scale was used with the HHX, written EFL exam with more specific descriptors relating to specific assessment criteria (See Appendix B5 for the specific scale descriptors). The assessment criteria are (see Appendix B4 for the Danish version of these criteria):

- Amount/Length (of text)
- Organizational Structure
- Ideas/Content and Use of Source Materials
- Language
- Style and Format

[25] The use of the same scale for all subjects throughout the educational systems facilitates calculating average scores and the possibility of including into the rater corps raters who are not teachers, but nevertheless are familiar with the scale.

[26] Originally, it was a 9 + 1 scale (13, 11, 10, 9, 8, 7 ,6, 5, 3 + 0) with the lowest score (0) indicating a misplaced student, i.e., a score indicating that the student cannot be assessed because the student's abilities are not adequate to be assessed at the particular level. It was also a score assigned to students who did not sit for the exam. The purpose of not having the levels 12 and 2 was to prevent the raters from giving the extreme scores of 13 and 3 (Undervisningsministeriet, 2004). As pointed out by the raters participating in this study and confirmed by Undervisningsministeriet (2004) it is often difficult to distinguish between the extreme scores (3 and 13) and the near-extreme scores (5 and 11) respectively, especially at the lower end of the scale. The distance between say a 3 and a 5 then is now considered a one-level difference. A new scale was introduced a year ago, i.e. after the collection of my data.

The weight of these official criteria in the holistic assessment is not stated or prescribed.

3.2.8 Appeal Process

When the students receive their final exam scores, they can appeal their scores (see Uddannelsesstyrelsens Internetpublikationer, 2001:8). In such appeal cases the raters who have assigned the appealed scores have to send written arguments for their scores to the school the student has attended. The student then has a week to comment on the raters' arguments. On the basis of an evaluation of the raters' arguments and the students' comments, the school (usually the principal) can decide to take one of the following actions:

- Re-rating (two new raters are appointed to rate the student's exam paper),
- Re-examination (the student sits for a new exam), or
- Denial: The school decides that the original score is maintained.

Appeals are rare (Karsten Sielemann, 2004 – personal communication), and if they do occur, it is usually with scripts that have received a failed score as their final score.

3.3 The Current Study

With assistance from the chair of the national rater corps and financial support from The Ministry of Education[27] I was able to carry out a study with 20 current members of the national rater corps and with authentic student scripts. As a result, the rating procedures were authentic with respect to the actual procedures and raters in an official HHX exam.

3.3.1 Scripts

The particular HHX EFL exam selected for this study was the exit exam of 2004[28] (see Appendix A). The focus of my study was part III (the composition) of this HHX EFL exam. The reason for

[27] Financial support was obtained through the Ministry's fund for research and development projects (Forsøgs- og Udviklingsmidler)

[28] This exam was selected because it was the most recent one when the data were to be collected.

choosing the composition task was that it gave students the opportunity to demonstrate their abilities to individually derive relevant information (from the source reading packet) and structure a coherent text set in a specific communication situation. Further, out of the three parts of the HHX EFL exam, Part III is the task that most closely resembles, and thus serves as a point of comparison to, the writing tasks in other, previous studies on rating behavior in ESL/EFL writing assessment.

In the 2004 exam the students were asked to write an essay answering the following question: "Skriv et essay om etik i forbindelse med produktion af varer og tjenesteydelser i tredieverdenslande – Skriv essayet på engelsk, og giv det en passende overskrift ... I besvarelsen inddrages oplysninger fra det engelske og det danske tekstmateriale" (*Write an essay on ethics in relation to the production of goods and services in the third world - Write the essay in English and give it an appropriate title ... include in your response information from the English and the Danish source materials* [my translation]). The source reading passages were an English newspaper article titled "Sewing a seam of worker democracy" and a Danish article titled "Reebok fører etiske korstog" (*Reebok leads the ethical crusade* [my translation]). Part of the students' task in this exam was to summarize the English text (Part I of this exam, which counts 25% towards the final score) and to translate the Danish text into English (Part II of this exam, which counts 25% towards the final score). These two subtasks were not included in the current thesis study, as the focus was on Part III – the composition, which counts 50% towards the final score (See Appendix A for the exam packet used in this study).

A total of 15 student essay scripts were selected for this study. All scripts were written by Danish learners of English, sitting for or preparing to sit for the written EFL HHX exam. It was not possible to obtain samples from the actual HHX exam exclusively due to logistical problems of getting in contact with students to obtain their consent to use their papers. The schools in charge of the HHX exams lose track of their students after the exams are completed, so although administrators at individual schools gave me permission to consult their former students' exam scripts, I managed to track down only 6 students who had sat for the actual exam. Thus, 6 out of the 15 student scripts used in this study were written under actual exam situations. The remaining

9 scripts were HHX mock exam papers written by students practicing for the test (i.e., taking classes leading up to the test). These mock exam papers were provided by the chair of the national rater corps.

My attempt was to collect student samples that represented the whole range of the scale from its high, middle, and low points, but for logistical reasons, I had to take what I could get. Although the chair of the national rater corps, who provided me with the mock exam papers, predicted that these papers would represent both ends of the scale, it turned out that the raters in this study scored relatively harshly[29]. So I ended up with scripts representing mostly the middle and the low end of the scale, with only a couple of samples representing the higher end of the scale. In the present study none of the students received the highest score, none received the second highest score, and 5% received the third highest score... and none received the lowest score. This distribution, in fact, parallels the scores in the real exam situation of 2004, in which 0% of the students received the highest score, 1% received the second highest score, 8% received the third highest score, and 0% received the lowest score (www.uvm.dk).

All the scripts were typewritten by the students originally, and they were for this study blinded as to the names or other identifying characteristics of the students[30].

The order in which the raters were presented with the student scripts was random[31], but not mixed for each rater[32]. As in real exam situations, the raters in my study rated the scripts in whatever order they chose[33].

---

[29] This trend to rate harshly in research settings echoes Lumley's (2005) study, in which he found that his raters scored more harshly in research settings than in operational settings (actual exam situations). Lumley explained this difference in harshness in terms of consideration for possible test takers: in a research setting, there will not be any adverse consequences for the test takers. Similarly, Broad (2003:79) found that raters in norming sessions were "stricter" or "harsher" because the scores had no real life consequences for the test-takers.

[30] Unlike Lumley (2002, 2005) and Johnson et al 2005, I did not specifically select writing samples that had been assigned discrepant scores in prior assessments. This might have generated lengthy discussions by the raters, and so been useful for the research. My main reason for not choosing this methodology, however, was to preserve the natural contexts for CWA rather than to focus exclusively on rater agreement (as in Lumley, 2002, 2005; Johnson et al., 2005).

[31] The scripts were shuffled to mix them up.

[32] Many researchers have chosen to mix the order of essays to prevent a contrast effect (Daly & Dickson-Markman, 1982), also known as a carryover effect. However, as in DeRemer (1998), I did not use this procedure out of a concern for authenticity: In the real exam situation, co-raters all receive the scripts in the same order.

3.3.2 Participants

Through a fund for research and development projects (Forsøgs- og Udviklingsmidler) administered by the Ministry of Education, it was possible to pay 20 raters from the national rater corps for HHX English to participate in the study[34]. All raters from the national rater corps (89 raters in total) were approached and coincidentally 20 raters volunteered (exactly the number of participants made possible by the fund from the Ministry of Education). All participants signed forms to indicate their informed consent in view of ethical considerations in the research. To maintain the confidentiality of the individual raters, I created pseudonyms for each rater to mask their identities but to preserve their genders.

According to the information the raters provided in the profile questionnaire (see Appendix C for questionnaire and summary of profiles) the 20 participants had the aggregate demographic profiles displayed in Table 3.1.

Table 3.1: Raters' Profiles

Gender	**Age**	**Teaching Experience**	**CWA Rating Experience**	**Highest Educational Level**
Male: 6 Female: 14	31-40: 3 41-50: 6 51-60: 7 >60: 4	Mean: 20.9 years Minimum: 7 years Maximum: 35 years	Mean: 15.3 years Minimum: 3 years Maximum: 31 years	MA: 18 BA: 2

Table 3.1 shows that there were more than twice as many female raters as male raters; they had a wide range of ages, with a concentration in the forties and fifties; and the vast majority had a Master's degree. These demographics largely correspond to the national rater corps as a whole (Karsten Sielemann, personal conversation – 2004). The raters were highly experienced and appropriately qualified, as should be the case for their admission to the national rater corps. Their

[33] Four out of 20 raters did, in fact, change the order of the essays, all of whom decided to read the shorter scripts before the longer ones.
[34] The stipends the raters received were equivalent to what they are regularly paid for rating student scripts.

expertise was determined by these external criteria and their membership in the national rater corps rather than by reliability checks, as in studies such as Wolfe et al. (1998)[35].

3.3.3 Procedure

To my knowledge, no other studies have sought to chronicle raters' decision-making behaviors in CWA practice from independent to communal ratings, so there was no well-established method to apply to trace rating behaviors of the raters in the present study. For this reason, I applied some of the methods of data collection established in certain previous studies of raters' decision-making when they rate independently in traditional writing assessments, and I drew on methods employed in the few studies conducted on communal ratings. These methods involved profile questionnaires, reflective questionnaires, think-aloud protocols, audio-recordings of discussions between raters, and analyses of scores assigned.

3.3.3.1 Data Sources

There were five data sources for the present research:

1. Profile questionnaires;
2. Audio recordings of concurrent think-aloud protocols (in independent rating sessions);
3. Audio recordings of communal rating dialogues (of communal rating sessions);
4. Retrospective questionnaires (raters' reflections on their own CWA practices); and
5. Scores assigned to the student scripts.

[35] Johnson et al. (2005:129) claimed to not select their raters on the basis of reliability, but rather "based on their experience and accuracy ". However, when determining the accuracy of their raters they calculated their "level of agreement with the scores assigned by the validation committee to exemplar papers", and the members of this validation committee were selected based on their high inter-rater reliability scores. Thus Johnson et al. (2005) can be said to have used reliability as a yardstick for measuring raters' accuracy.

3.3.3.1.1 Concurrent Think-Aloud Protocols (TA) – Independent Ratings

Rater participants were asked to produce audio-taped, concurrent think-aloud protocols while rating the student scripts during their independent rating sessions. That is, they rated the scripts as they would in regular exam settings, but with the added task of thinking aloud to document what they paid attention to while they performed their ratings. The think-aloud protocols were intended to shed light on the cognitive processes the raters went through when rating, in particular the factors influencing the raters' decisions as well as how they monitored themselves when interpreting and judging students' scripts.

Numerous researchers have employed TA methods to document raters' independent, decision-making behaviors during traditional writing assessment practices (e.g., Cumming, 1990; Cumming et al. 2001, 2002; DeRemer, 1998; Erdosy, 2004; Lumley, 2002, 2005; Milanovic et al., 1996; Pula & Huot, 1993; Sakyi, 2000, 2003; Vaughan, 1991; Wolfe, 1997; Wolfe et al., 1998). Generally, TA has become a respected research tool for tracing human cognitive processes in a substantial number of studies related to education and psychology (e.g., Hayes & Flower, 1983; Perl, 1979; Raimes, 1985; Swarts, Flower, & Hayes, 1984). The claims for TA as a valid method of inquiry into human thought processes have largely been based on the work of Ericsson and Simon (1984/1993), who presented a well-developed theory supported by elaborately reviewed empirical evidence. The core of their model is the assumption of an explicit relationship between the contents of short-term memory and verbal reports. Ericsson and Simon claimed that "the information that is heeded during the performance of a task is the information that is reportable; and the information that is reported is the information that is heeded" (1993:167). The general idea is that people producing TAs report on information that they attend to during the performance of a task. The reports are assumed to be mostly unaffected and unedited by selective evaluations and inferences and thus are relatively pure records of data about their thinking processes[36].

[36] The assumption that concurrent protocols prevent raters from editing their impressions of the scripts they are rating casts doubt on the validity of exclusive use of retrospective reports to document rater behavior, even though retrospective reports have been used in several studies of rater behavior (e.g. Connor-Linton, 1995a; Mendelsohn & Cumming, 1987; Shi, 2001; Song & Caruso, 1996).

Although many researchers have accepted TA as an unproblematic method of obtaining data on writing assessments (e.g., DeRemer, 1998, Huot, 1993; Vaughan, 1991), a number of researchers (notably, Lumley 2005; Stratman & Hamp-Lyons, 1994) have cautioned against assuming a direct relationship between what is reported in verbal protocols and the cognitive processes participants undergo. The controversial issues around the validity of TA pertain mainly to the notion that the verbal reports may be incomplete and/or altered by the artificial setting surrounding the procedures:

1. Incomplete reports: Much of what goes on in people's minds is automatic and/or disorganized, so it is not all reportable. Many cognitive processes are inaccessible, so researchers must, when analyzing TA protocols, accept that the verbal reports do not provide comprehensive records or full insights into cognitive processes.
2. Alteration: The performance reported during TAs may not reflect the performance during real, natural, or operational settings. The performance may be altered by the artificiality of a research setting or awareness of an observer or that analyses will later be conducted. Further, the additional task posed on the participants to report unorganized thoughts in an organized and linear manner may alter the process itself.

The validity of verbal reports in writing assessment are further compromised because assessing writing is an ill-structured task (Bracewell, & Breuleux, 1994; DeRemer, 1998): Raters don't entirely agree on what contributes good writing, making it practically impossible to conduct controlled experiments to determine whether verbal reports do, indeed, reflect such a reality. One study, Barkaoui (2007a), did examine the effects of TA on raters' decision-making behaviors. Although Barkaoui (2007a) found some changes in rater behavior caused by the TA requirement (e.g., raters becoming slightly more severe, rater internal consistency shifting from overfit to acceptable fit to misfit, and certain changes in rating criteria), these changes were either non-significant or just barely significant in terms of altering the behavior of raters when they were not producing TAs.

Scholars such as Dechert (1987), Grotjahn (1987), Lumley (2005), Matsumoto (1993), and Smagorinsky (1994a, 1994b) have concluded that there is value in TA methods. There is no reason to assume that what is reported during TAs is not heeded by research participants (Matsumoto, 1993:50). As Smagorinsky (1994a) argued, certain things about writing or assessing can only be revealed by TAs. So it seems that although TA may not provide a comprehensive account of the human cognitive processes of writing assessment behaviors, "our position should...be one of accepting the partial nature of the information we receive" (Lumley 2005:77). In acknowledging the limitations of TA, however, researchers need to try to diminish those limitations, to be cautious of claims made, and to accept that such data can only "be considered as indicators of the cognitive processes, not direct evidence of their full realizations" (Cumming: 1990:43).

3.3.3.1.1.1 Instructions

Analysts of TA methods (e.g., Green, 1998; Matsumoto, 1993; Pressley & Afflerbach, 1995; and in particular, Ericsson & Simon, 1984/1993) have argued that it is possible to diminish the limitations of TA by giving careful consideration to the research setting and the instructions that participants receive. In planning the data gathering settings and introducing my participants to the TA procedure, I paid special attention to the above mentioned threats to the validity of the use of TA: the potential alteration and incompletion of the verbal reports.

To get the participants in my study to verbalize as much of their cognitive processes as possible (and thereby reducing the incomplete nature of the protocols), they were trained and warmed-up individually according to uniform instructions following Ericsson and Simon's (1984/1993) guidelines: I oriented them to the TA procedure with arithmetic problems and then gave oral instructions to rate the set of scripts as they would for a real exam, but to think aloud while doing this (see Appendix D for warm-up exercises). Following Ericsson and Simon's (1984/1993) guidelines, the participants in this study were asked not to explain or analyze their thoughts, but merely to vocalize what naturally passes through their heads while rating. I conducted the training face-to-face with each participant.

To keep up a steady stream of speech the raters were given a note with bold letters saying "tænk højt" ("*think aloud*"). My presence was offered during the TA task, but rejected by all participants. They expressed that after the TA warm-up they felt comfortable with the procedure, but more importantly: they felt that the presence of the researcher might affect their performance because they usually conduct the independent ratings in private and in surroundings chosen by themselves. That none of the participants in my study opted for my presence to assist and/or prompt them in their verbal reports, and instead chose their usual rating surroundings, reduces the artificiality of the verbal reports and thus diminishes their potential alteration during the research setting. This limited researcher intervention in TA is in line with Pressley and Afflerbach, who contend that "given the frequent observation in the social sciences literature that people will often comply with researcher demands, we have to conclude that researcher silence about how the script might be processed is more defensible than directions that prompt particular processes, especially when the goal is to learn about the processes people naturally use when they read" (1995:132-3).

The artificiality of the research setting (and consequently the risk that the raters' verbal reports might not reflect the cognitive processes of an operational, non-research, setting) was further diminished by participants rating in settings that were familiar to them, and so they did not have to be trained for the specific rating task: They were presented with essay scripts they had many years of experience teaching and with a rating context which they were used to (i.e., a familiar rubric, the CWA procedure). In most other research studies of rating behavior, raters were trained to rate under rating conditions they were not entirely familiar with, for example, rating with no rubric at all (e.g., Cumming et al., 2001, 2002) or using rubrics they had no or little experience in using, and thus which they had to be specifically trained in using (e.g., Lumley, 2002, 2005). By avoiding training raters for the specific rating task I created a research situation that was non-intervening and observational in character (as in Sakyi, 2003). Not imposing experimental-type control limited the possibility of my unintentionally restricting or mandating the raters' behaviors and thus corrupting the results from the verbal reports. However, despite my attempts to create an environment and a mindset akin to that of an operational rating setting,

there was still the risk that the raters may have been affected by the purpose of the research and this setting: The purpose of their rating in the research setting was for me to conduct research, whereas in an operational setting, the purpose would be for the students to receive scores in a real-life, high-stakes test.

3.3.3.1.2 Audio Recordings of Co-Rater Dialogues in Communal Ratings

To continue tracing the raters' decision-making behaviors into the communal rating sessions, I employed an observational method of (audio) recording the co-rater dialogues. Concurrent think-aloud protocols would not be a suitable data-gathering method because they would interfere with the natural flow of the dialogues. Also, what was of interest in the CWA sessions was not so much what went on in the minds of the raters, but what they brought up in the discussions: what (and how) the raters choose to put forth in justifying their perceptions and in challenging the views of their co-raters to reach a final score. Audio recordings of rater dialogues have been employed in other research into communal rating behavior in writing assessment (e.g., Allen, 1995[37]; Broad, 1997; Broad, 2000; Broad, 2003; Mohan & Low, 1995).

3.3.3.1.2.1 Instructions

There was no researcher intervention prior to or during the recordings of the communal rating sessions: The participants were requested to enter the sessions as they would in a normal exam situation, the only difference being the cassette recorder placed between the two raters. As with the independent rating sessions, I was not present. Also, no training was necessary because all the participants were familiar with and highly experienced in the communal rating practices related to the scripts being rated.

3.3.3.2 Analysis of Verbal Data

[37] Allen (1995) did not record his observations on audio cassette, but he kept records of email exchanges between raters.

The TA protocols obtained from the independent rating sessions and the recordings of the rater discussions from the communal rating sessions were analyzed to answer the specific research question of raters'disctinct decision-making behavior and the sequence of these behaviors in the two rating sessions. The raters' textual focus was consulted to address the research question of the extent to which the raters adhered to the official assessment criteria. The verbal reports in the communal rating sessions were further analyzed to address the research question of equal engagement between the raters.

The analyses were conducted largely according to procedures suggested by Green (1998) and Miles and Huberman (1984/1994) and practiced by Cumming et al. (2001) and Cumming et al. (2002). The methods of transcription, segmentation of verbal reports, and coding were as follows.

3.3.3.2.1 Transcriptions

The verbal data from the independent rating sessions and the communal rating sessions were recorded and transcribed in full. These recordings totaled 600 protocols: 300 for the independent rating sessions (20 raters rating 15 scripts each) and 600 for the communal rating sessions (20 raters (10 rater dyads) rating 15 scripts), totaling 1,964 pages (1,548 for the independent ratings and 416 for the communal ratings) and 355,689 words (289,863 for the independent ratings and 65,826 for the communal ratings). See Appendix E for transcription conventions.

3.3.3.2.2 Segmentation, Coding, and Preliminary Analysis

To put the data in a manageable, quantifiable form and to obtain an overall picture of the trends and patterns in the raters' decision-making behaviors, all the verbal transcripts were segmented, coded, and analyzed in full. As became clear in the process of segmenting the data and designing the coding scheme, a strong element of analysis was involved in segmenting and especially in coding. Thus the segmentation phase, the coding phase, and the analysis came to form an iterative and interpretive process, subject to continual modification, as is typical in qualitative data analysis (Green, 1998; Miles & Huberman, 1884/1994).

3.3.3.2.3 Segmentation

Segmentation is a way of dividing the protocols into manageable and consistent units for the purpose of analysis. There are many different ways to divide verbal data. Green (1998) suggested that each segment of verbal data represent a different process and that it primarily comprises a phrase, a clause, or a sentence. In segmenting his data, Lumley (2005) adopted a more pragmatic view of segmentation, recognizing Paltridge's (1994) contention that textual boundaries are made according to the content of the script rather than exclusively along linguistic boundaries. Cumming et al. (2001, 2001) and Cumming (1990), too, segmented their protocols largely according to content as well as utterance boundaries. Adapting a coding scheme from Cumming et al. (2001, 2002), I chose to follow their guidelines for segmentation. They use the following three criteria in dividing their protocols into separate, comparable units:

Pauses of five or more seconds (marked by three dots (...) in the transcriptions),
The reading of a segment of the student script (marked by capital letters in the transcriptions),
The start or the end of an assessment.
(Cumming et al. (2001:34)

To maintain units that could be comparable across the two rating sessions, I attempted to apply the segmentation criteria used for the independent rating sessions to the communal rating sessions. The three criteria for segmenting the independent rating sessions were maintained, but the following criterion was (initially) added: Each conversational turn signaled a new segment[38].

I quickly realized that in order to allow for characteristics of natural conversations such as interruptions and conversational gambits, this criterion had to be modified or at least defined more precisely. Thus the modified criterion came to be:

[38] See Allwright and Bailey (1991) for methods of segmenting conversational data by conversational turns.

Each conversational turn signaled a new segment if that turn was not an uptaker such as 'aha', 'I agree', or a mere repetition of what the other person just said and with no new information added.

3.3.3.2.4 Coding and Preliminary Analysis

I coded all of the segmented verbal report data from the independent rating sessions and the communal rating sessions. The coding scheme had to address the specific research questions related to the raters' decision-making behaviors while adequately describing the data I had collected for this purpose. Like Cumming et al. (2001:34) I considered each verbal segment to involve at least one, but potentially several, decision-making behaviors. These were all coded according to a coding scheme, which came to be a combination of an a priori model or conceptual framework and my interpretive attempts to let the data speak for themselves. As all the transcripts were comprehensively being analyzed, the coding scheme underwent several revision and refinement processes until it could adequately account for the data. The final coding scheme was applied to systematically document the trends and patterns of the raters' decision-making behaviors in both the independent rating sessions and the communal rating sessions, and it facilitated a comparison between the raters' behaviors in the two rating sessions.

3.3.3.2.5 Coding Scheme: Conceptual Framework

The conceptual framework (Miles & Huberman, 1984/1994) of the coding scheme was drawn from previous studies into raters' distinct decision-making behaviors in traditional performance-based writing assessment. In particular, the framework was inspired by Cumming et al. (2001), Cumming et al. (2002), and Cumming (1990), who applied a coding scheme which had been empirically developed and rigorously validated. The main dimensions in their coding scheme, and in mine, are: (a) strategies: Interpretation Strategies and Judgment Strategies and, cutting across these strategies, (b) the focus of rating, either a Monitoring or Contextual focus or a Textual Focus. These dimensions are displayed in the Table 3.2 and elaborated upon below.

Table 3.2: Major Categories in Coding Scheme

	Monitoring or Contextual Focus	Textual Focus
Interpretation Strategies		
Judgment Strategies		

3.3.3.2.5.1 Interpretation Strategies and Judgment Strategies

As described previously in Chapter 2, Interpretation Strategies represent strategies raters use to construct an image of the scripts they are rating, whereas Judgment Strategies signify strategies raters employ to evaluate their constructed images of the scripts. The distinction between interpretation strategies and judgment strategies has its origin in Freedman and Calfee's (1983) model of rater behavior, which proposed that raters create an image of the script they are rating and subsequently evaluate and judge that image. Many researchers have acknowledged their proposition that raters don't evaluate the scripts themselves, but rather the image they have created of the scripts. This has generated an interest in, on the one hand, how raters construct or interpret the scripts to be rated and, on the other, how they evaluate and judge these images of the scripts. This is reflected in coding schemes applied to investigate rater behaviors, such as Cumming et al. (2001, 2002), Cumming (1990), and Erdosy (2004), who all specifically distinguished between Interpretation Strategies and Judgment Strategies. Other researchers, too, have made this distinction between interpretation strategies and judgment strategies in their coding schemes (even if those exact terms have not been applied or have been split into minor strategies). Among those are Wolfe (1997) and Lumley (2002, 2005). Wolfe (1997) distinguished between Interpretation Strategies, Evaluation Strategies, and Justification Strategies (the latter two strategies largely making up what Cumming et al. (2001, 2002), Cumming (1990), and Erdosy (2004) would call Judgment Strategies). Lumley (2002, 2005) made a distinction between Reading Behaviors and Rating Behaviors, which roughly correspond to the present distinction between Interpretation Strategies and Judgment Strategies, respectively.

(Lumley also used a category he termed Management Behaviors, which corresponds to Cumming et al.'s, (2001, 2002), Cumming's (1990) and Erdosy's (2004) Monitoring Focus).

3.3.3.2.5.2 Rater Focus (Textual Focus and Monitoring or Contextual Focus)

Cutting across the Interpretation Strategies and Judgment Strategies is the dimension of Rater Focus, which represents what the raters focused their attention on when constructing an image of the scripts and when judging that image. This, again, corresponds to Cumming et al.'s (2001) and Cumming et al.'s (2002), Cumming's (1990), and Erdosy's (2004) Rater Focus, although I decided to define the Rater Focus in the present study slightly differently and, in more detail, to investigate the extent to which the raters seemed to adhere to the official assessment criteria prescribed for the HXX exam. The major rater focus categories are:

A) Monitoring or Contextual focus, which signifies a focus on how the raters monitor their own rating behavior or that of their co-rater, including considerations of the context. This focus applied when raters interpreted the scripts as well as when they judged them. In addition to the coding scheme of Cumming (1990), Cumming et al. (2001) and Cumming et al. (2002) some form of a monitoring focus has also been included in various other coding schemes describing rater behavior. It is akin to Lumley's (2002, 2005) Managing Behaviors and his focus on external factors in his Rating Behavior, and to Wolfe et al.'s (1998) Processing Action categories, and it bears some resemblance to DeRemer's (1998) Processes or Operations and to Heller, Sheingold and Myford's (1998) Rating Process actions.

B) Textual Focus, which indicates what textual features the raters focus on when interpreting the scripts and when judging them. The textual foci in the coding scheme reflect the official assessment criteria and text characteristics: Amount of Text, Organizational Structure, Content (ideas) and Use of Source Materials, Language, and Style and Format. With the exception of Amount of Text, these categories reflect those used in other coding schemes (again with slightly different terms), such as Cohesion and Organization, Task Fulfillment and Appropriacy, Grammatical Control, Conventions of Presentation in Lumley's (2002, 2005) study;

Organization, Storytelling, Mechanics, and Style in Wolfe et al.'s (1998) study[39]. In specifying the categories of Content and Language, I constructed a detailed coding scheme to establish a clear account of the extent to which the official scoring criteria were reflected in the raters' comments. As Green (1998) argued, the flipside of such a detailed scheme, however, is that inter-coder reliability is usually sacrificed (as described below).

The subcategories of the coding scheme emerged from my thorough review of the data set and were finalized after repeated modifications and the co-coder work. The final coding scheme with specific subcategories included in the two dimensions is reproduced in Table 3.3 (See Appendix F for samples of coded data).

Many of the specific categories bear close resemblance to the categories in Cumming et al.'s (2001) and Cumming et al.'s (2002) coding scheme. However, I added certain specific categories to accommodate the present research focus and data. Not only did I expand the range of textual focus to reflect the extent to which the raters adhered to the official assessment criteria, but I added more Judgment Strategies to account for the complex dynamics that evolved during the communal ratings. Because research on communal rating behaviors is still in its infancy, no sound theoretical frameworks or validated coding schemes could be drawn upon in the construction of the specific coding categories. The code specifications here are thus exploratory and linked to the present data, resembling the methods of empirically grounded coding (Glaser, 1978). I had initially intended to include categories that reflected how and to what extent the raters applied the official scoring scale as in Lumley's (2002, 2005) study. However, it became apparent during the coding of the data that, despite their respecting the official assessment criteria, the raters never consulted or directly applied the scale or the scale descriptors, and thus, to be faithful to the present data, I did not include such a code in the coding scheme[40].

[39] Unlike Lumley (2002, 2005) and Wolfe et al. (1998), and the present study, Cumming et al. (2001, 2002), Cumming (1990), and Erdosy (2004), did not investigate to what extent raters adhered to the official assessment criteria, and so did not feel compelled to specify the categories of Content and Language further.

[40] This is, in fact, a complex matter, as one cannot conclude that the raters ignore the scales and the descriptors, as they may have internalized the scale or just been so familiar with the scale that they did not bother to mention it explicitly.

Table 3.3 Final Coding Scheme

	Contextual or Monitoring Focus CM	**Textual Focus T**				
Interpre-tation strategies I	**I-CM**	**IT-V Amount**	**IT-STR Organiza-tional Structure**	**IT-C Content and Use of Source Materials**	**IT-L Language**	**IT-STY Style and Format**
	1. Read or interpret task input/source material 2. Read or reread student script 3. Envision personal situation of the student 4. Consider task or exam require-ments 5. Consider own perception of correct English (e.g. consult a dictionary)	1. Scan for length	1. Discern or scan for organizational structure	1. Discern or summarize ideas 2. Identify or interpret ambiguous or unclear phrases	1. Classify language errors into types 2. Identify errors 3. Edit or correct language (errors or unclear phrases)	1. Discern style, register or genre
	Contextual or Monitoring Focus CM	**Textual Focus *T***				

Judgment strategies J	**J-CM**	**JT-V Amount**	**JT-STR Organiza-tional Structure**	**JT-C Content and Use of Source Materials**	**JT-L Language**	**JT-STL Style and Format**
	1. Articulate score 2. Compare student scripts 3. Define, revise or suggest assessment strategies 4. Articulate general impression 5. Deliberate/articulate teaching practices 6. Exemplify directly from student script 7. Consider consensus-based strategy 8. Consider personal response or bias	1. Assess or justify amount of text	1. Assess organizational structure overall 2. Assess or justify title 3. Assess or justify introduction and/or conclusion 4. Assess or justify coherence and/or cohesion	1. Assess content/ideas overall 2. Assess or justify reasoning, logic, or topic development 3. Assess or justify clarity 4. Assess or justify correctness of or disagreement with content 5. Assess or justify maturity or independence 6. Assess or justify task/topic relevance or completion 7. Assess or justify use and understand-ing of source material	1. Assess or justify language overall 2. Assess or justify frequency of errors 3. Assess or justify gravity of errors 4. Assess or justify syntax or morphology 5. Assess or justify lexis 6. Assess or justify fluency or comprehen-sibility 7. Assess or justify spelling 8. Assess or justify punctuation	1. Assess or justify style 2. Assess or justify genre

One might question my use of the same coding scheme with data elicited under different conditions (i.e., recordings of think-aloud protocols vs. recordings of dialogues). However, in order to trace the rating behaviors throughout the rating procedure from independent rating sessions to communal rating sessions a scheme was needed to facilitate comparative analyses across the two sessions in a coherent manner. Undoubtedly, the number of interpretation strategies significantly decreased in the communal ratings, compared to the independent ratings, because the raters had already interpreted the scripts to be rated before they entered the communal rating sessions.

To faithfully account for the data, the coding scheme underwent several refinements and moderations, in which codes were eliminated, merged, split up, added, and moderated. The final adjustments were made after the co-coder work. However, despite a recursive approach to continuously improve the coding scheme, it was impossible to keep all the categories entirely discrete, as also experienced by Cumming et al. (2001, 2002) during their analyses. The raters themselves noticed and commented on this interrelatedness, e.g., "kommer igen noget der er uklart formuleret, om det er det engelske eller om det er uklart" (*Susanne, script 3)*; *translation: "again something unclear, whether it is because of the language or because it is really unclear"*). In assessing such an ill-defined task as writing, raters are dealing with underlying constructs that are not easily separated or definable, and the fact that the raters themselves commented on the overlaps adds to the authenticity and the fuzzy, interdependent quality of the categories in the coding scheme. What further added to the challenge of coding was that the codes were designed to reflect not the textual aspects as such, but the raters' perceptions of those aspects: Raters may not only perceive the scripts differently but they may use different language to frame the same textual aspects (as also experienced by Heller, Sheingold & Myford, 1998).

3.3.3.2.6 Inter-Coder Reliability

My research purposes and the fuzzy boundaries of the codes made it imperative that the coding scheme be subjected to inter-coder reliability checks. I first conducted an intra-coder reliability check on a sample (10% of total protocols) of my own coding, and when my intra-coder

reliability rate reached an acceptable level (89%), I subjected the coding scheme to an inter-coder reliability check. A recent linguistics graduate from Georgetown University, Washington, D.C. volunteered to work with me to reach an acceptable inter-coder reliability level. I trained the second coder with my operational definitions of the categories in the coding scheme along with some illustrative exemplar samples from the protocols to anchor each coding category. (These samples translated into English are shown in Appendix F.) With the use of these guidelines the second coder and I individually coded a small number of protocols. We then worked together to resolve discrepancies in our perceptions of what the different coding categories represented. After a few recursive rounds of refining the categories, we subjected our codings to an inter-coder reliability check and established an average inter-coder reliability rate of 84% (percentage agreement) on 60 (=10%) randomly selected protocols. This level of agreement suggested that the coding scheme could be applied with adequate reliability. I decided that an acceptable or adequate level of reliability would be a level comparable to the average reliability levels reported in other writing assessment studies using verbal protocols. These levels have been reported to above 80%. For instance, Cumming et al. (2001, 2002), using percentage agreement, reported an average inter-coder reliability level of 84%; Cumming (1990), also using percentage agreement, reported a reliability level of 87%; Erdosy (2004), also using percentage agreement, reported a reliability level of 84%; Wolfe et al. (1998), using Cohen's kappa reported levels between .85 and .93; DeRemer (1998) reached a reliability level of 93%; Lumley (2002, 2005) reported reliability levels of 94%. My inter-coder reliability level of 84% may be on the lower end of the reliability levels reported in comparable research. This is probably due to the high level of specificity of my coding scheme (see Green, 1998).

3.3.3.2.7 Coding Management

All 600 protocols were coded, the results of which were entered into Microsoft Excel spreadsheets. This is unlike the procedures taken by Cumming et al. (2001, 2002) or Lumley (2002, 2005). The former selected only those protocols that represented essays that were scored consistently by most raters, and the latter selected only those protocols that yielded inconsistent scores. The reason for my not screening the scripts in the present study for consistency or

inconsistency in scores is that I was interested in general rater behaviors in all types of rating situations, including ratings that yielded consistent as well as inconsistent scores.

All the coded data were converted to percentages of the total number of decisions made by individual raters for each script. This was done mainly to account for the fact that the raw counts varied greatly from rater to rater reflecting differences in their verbosity. Converting the data in this way made it amenable to comparative analyses. Cumming et al. (2001) reported that this procedure is common in analyses of think-aloud data using large numbers of coding categories (i.e., the coding scheme used in the present study consisted of 43 categories).

To represent the trends in raters' decision-making behaviors simple descriptive statistics (mean percentages and standard deviation and grand mean percentages for the aggregated behaviors) involving summary tables and charts were used.

3.3.3.2.8 Analysis of Verbal Data and Scores to Determine Level of Equal Engagement

To answer the specific research question of the level of equal engagement in communal rating sessions, the rater dyads were examined for score dominance and conversational dominance as well as the interaction between score dominance and conversational dominance. Score dominance refers to the distance between each of the independent scores and the final, communally rated score. If the rater whose independent score was closer to the final communal score than his/her co-rater was, then that rater dominated by score. If, however, his/her independent score was further away from the final score than his/her co-rater's score, then s/he conceded his/her score. This approach to determining the extent of equal engagement between raters was used by Johnson et al. (2005). Moss et al. (1998), however, focused on conversational dominance to determine the level of equal engagement between or among raters in a communal rating session, in that they did not look at scores but at how much each rater contributed to the conversations in the rater dyads (although there was no report in their article on exactly what features of the conversations were analyzed to determine the raters' level of contribution to the discussions).

Conversational dominance is a complex matter and can, of course, be measured in various ways, such as verbosity, initiating turns, interruptions, syntax structures, or choice of words (see Itakura, 2001 for various approaches to measuring conversational dominance). Well aware that a multiperspectival approach to operationalizing conversational dominance would offer a more in-depth analysis, I have chosen to take the quantitative measure of counting words and counting decision-making behaviors. Counting words offers an indication of verbosity. So would counting decision-making behaviors, but as uptakers and mere repetition of words did not count as independent decision-making behaviors, the number of decision-making behaviors would also add some degree of topic introduction and topic control (see Sections 3.3.3.2.3 and 3.3.3.2.4 on segmenting and coding the verbal data).

To get a more elaborate picture of whether the raters in communal rating sessions were "acquiescing to the more assertive voice" (Moss, 1996:26) and to see whether Johnson et al.'s (2005:126) assumption that score dominance is a product of conversational dominance makes sense, the relationship between score dominance and conversational dominance was examined. Each case of score dominance was examined for conversational dominance by checking whether the rater in the rater dyad who dominated by score would also be the one who dominated the conversation.

3.3.3.3 Retrospective Questionnaire

All raters were asked retrospectively (after finishing all CWA sessions) in a questionnaire about their perceptions of CWA (in general and in relation to the present rating sessions). Part 1 of the questionnaire asked which score, in hindsight, they found more accurate: their own independent score or the communally rated score; Part 2 asked what they thought about CWA procedures in general. (The questionnaire is shown in Appendix G.) The raters' responses served to shed light on how valid the raters themselves perceived CWA to be. In part 1 of the questionnaire the raters' perceptions of the accuracy of the scores were counted, and in part 2 of the questionnaire

their comments were analyzed impressionistically and are presented in the thesis by tendencies and illustrative comments.

3.3.3.4 Scores Assigned to Student Scripts

The scores assigned to the student scripts in the two rating sessions are an integral part of the raters' decision-making process and lend themselves to score range and rater agreement measures. As pointed out above, the 10-point scale used ranges from 0 to 13 with no score points of 12, 4, 2, or 1. To facilitate agreement and score range calculations, the scale was converted to a scale with no "missing" levels (see Appendix B2 for scale conversion).

3.3.3.4.1 Instructions

No constraints or specific instructions were given to the raters on how to assign scores for the purposes of this thesis research. They were requested to assign scores as they normally would in regular exam situations. This is unlike Johnson et al. (2005), who, in investigating the impact of discussion on reliability, specifically instructed their participants to "reach a consensus and to assign either of the original scores or to assign the average of the two scores" (2005:131). The reason for not giving the raters in the present study such constraints is that the raters would then be imposed scoring restrictions not reflective of real exam situations. In real exam situations the raters are allowed to assign any final score they find most accurate based on their communal rating discussions, whether the final score is one of the original scores, an average score, or neither. In fact, two raters in this study decided, when rating one of the scripts, that neither of their original, independent scores, nor an average score was appropriate, and they finally opted for a 'third' score, lower than both of the independent scores (Pernille and Jesper rating script 14 – see Appendix H for full range of scores).

3.3.3.4.2 Analysis of Scores

To determine the agreement level among raters in each of the two rating sessions, Kendall's Coefficient of Concordance (SPSS 13.0) and Cronbach's Alpha (SPSS 13.0) were used. Kendall's Coefficient of Concordance measures the degree of concordance among assessments and is calculated by rank ordering the judgments made by each judge (Reinholt Petersen, 2001). Cronbach's Alpha (coefficient alpha) estimates the internal consistency basing it on variances of the independent items (Bachman, 2004). Kendall's Coefficient of Concordance was used in Mendelsohn and Cumming (1987) to determine the agreement among multiple judges in two different sets of rating sessions and is recommended by Petersen (2001) to be used in nonparametric measurements of the degree of concordance among multiple assessments. Cronbach's alpha has also been used in writing assessment studies (e.g. Shi, 2001) focusing on inter-rater reliability rates and is referred to by the Norwegian Ministry of Education[41].

3.4 Summary of Methods

Empirical data were collected and analyzed to map raters' decision-making behaviors in CWA and chronicle the progression of these behaviors between independent rating sessions and communal rating sessions within CWA. The venue of the study was a well-established CWA practice in Denmark, and data were obtained from 20 experienced CWA raters rating the same 15 student scripts independently and with a co-rater. Their verbal reports were transcribed, segmented, coded and analyzed according to a coding scheme whose main categories were conceptualized and validated from earlier studies into raters' decision-making behaviors and whose specific categories emerged to account for the data of the current study. The final coding scheme recorded the raters' interpretation strategies and their judgment strategies as well as their textual and monitoring foci during their rating process and thereby served to address the research question related to mapping and comparing raters' distinct decision-making behaviors in their independent ratings and in their communal ratings, including how the raters distribute their attention to the official assessment criteria. To address the research question of equal engagement between raters in CWA discussions, the communal rating sessions were analyzed

[41] The Danish Ministry of Education has not issued inter-rater reliability guidelines, which is why I refer to a country whose educational system is comparable to the one in Denmark. The source is taken from www.utdanningsdirektoratet.no (Rammeverk for nasjonale prøver 2007).

for score dominance and conversational dominance as well as the interaction between score dominance and conversational dominance. To shed further light on the progression of rater behavior between the two rating sessions, score range and rater agreement levels were calculated and compared in the two rating sessions using Kendall's Coefficient of Concordance and Cronbach's Alpha. Finally, raters' perceptions of CWA were laid out by collecting their retrospective comments.

Chapter 4
Raters' Decision-Making Behaviors in Independent Rating Sessions

4.1 Purpose and Scope of the Chapter

This chapter reports on the findings of the raters' decision-making behaviors in the independent rating sessions, which is where the raters assigned their preliminary scores in preparation for the communal rating sessions. The chapter introduces the results by presenting the trends of the raters' distinct decision-making behaviors. I first present the raters' focus when they interpret the student scripts and then when they judge the scripts. To demonstrate the raters' decision-making tendencies, illustrative examples of rater behaviors are excerpted and placed in the context of frequencies of behavior (corresponding to the codes in the coding scheme), for as Miles and Huberman indicate "words and numbers keep one another analytically honest" (1984:55). As most of the coded categories of decision-making behaviors accounted only for a small portion of each raters'overall decision-making, a large part of the data is aggregated under larger categories. After mapping the raters' distinct decision-making behaviors, the prototypical sequences of their behaviors are laid out as well as the extent to which their textual focus corresponds to the official assessment criteria. Finally, the levels of rater agreement are reported.

4.2 Raters' Distinct Decision-Making Behaviors

This part of the findings presents the raters' distinct decision-making behaviors in the independent rating sessions. Table 4.1 shows the frequency counts of the raters' decision making behaviors identified during the independent rating sessions. The coded data were converted to percentages of the total number of decisions made by the individual raters for each script they rated.

Table 4.1: Mean Percentages (M) and Standard Deviations (SD) of Raters' Decision-Making Behaviors in Independent Rating Sessions

20 Raters and 15 student scripts = 300 protocols	***M(SD)***
Interpretation Strategies	
Contextual or Monitoring Focus	
Read or interpret task input/source materials	0.1% (0.4)
Read or reread student script	29.3% (12.2)
Envision personal situation of the student	0.8% (1.5)
Consider task or exam requirements	0.2% (0.6)
Consider own perception of correct English (e.g. consult a dictionary)	0.2% (0.8)
Textual Focus (Amount of text)	
Scan script for length	0.2% (0.7)
Textual Focus (Organizational Structure)	
Discern or scan for organizational structure	1.2% (2.0)
Textual Focus (Content and Use of Source Materials)	
Discern or summarize ideas	1.8% (2.6)
Identify or interpret ambiguous or unclear phrases	2.8% (3.1)
Textual Focus (Language)	
Classify language errors into types	7.9% (7.8)
Identify errors	4.8% (5.2)
Correct or edit language (errors or unclear phrases)	13.4% (9.6)
Textual Focus (Style and Format)	
Discern style, register or genre	0.7% (1.5)
Judgment Strategies	
Contextual or Monitoring Focus	
Articulate score	3.1% (2.3)
Compare student script	0.7% (1.5)
Define, revise or suggest assessment strategies	0.4%(1.5)
Articulate general impression	1.8% (2.8)
Deliberate/articulate teaching strategies/practices	0.1% (0.4)
Exemplify directly from student script	1.4% (5.3)
Consider consensus-based strategy	0.0% (0.3)
Consider personal response or bias	0.0% (0.0)
Textual Focus (Amount of Text)	
Assess or justify amount of text	1.8% (2.3)
Textual Focus (Organizational Structure)	
Assess or justify organizational structure overall	0.7% (1.4)
Assess or justify title	1.2% (1.7)
Assess or justify introduction and/or conclusion	1.3% (1.8)
Assess or justify coherence and/or cohesion	0.8% (1.6)
Textual Focus (Content and Use of Source Materials)	
Assess or justify content/ideas overall	1.2% (1.9)

Assess or justify reasoning, logic, or topic development	1.6% (2.4)
Assess or justify clarity	0.3% (0.8)
Assess or justify correctness of or disagreement with content	1.5% (2.5)
Assess or justify maturity or independence	0.9% (1.8)
Assess or justify task/topic relevance or completion	2.0% (2.9)
Assess or justify use and understanding of source material	1.7% (2.6)
Textual Focus (Language)	
Assess or justify language overall	2.6% (2.9)
Assess or justify frequency of errors	2.1% (2.4)
Assess or justify gravity of errors	0.8% (1.6)
Assess or justify syntax or morphology	2.5% (3.6)
Assess or justify lexis	1.4% (2.4)
Assess or justify fluency or comprehensibility	3.4% (3.2)
Assess or justify spelling	0.4% (1.2)
Assess or justify punctuation	0.3% (0.9)
Textual Focus (Style and Format)	
Assess or justify style	0.5% (1.3)
Assess or justify genre	0.4% (1.2)

For an overview of the decision-making tendencies, Table 4.2 provides the same findings as Table 4.1, but in an aggregated form corresponding to the main categories of the coding scheme.

Table 4.2: Grand Mean Percentages (M) and Standard Deviations (SD) of Raters' Decision-Making Behaviors in Independent Rating Sessions

20 *raters rating 15 scripts each=300 protocols*	Conxtext ual or Monitor-ing Focus	Textual Focus					Total
		Amount of Text	Organiza-tional Structure	Content and Use of Source Material	Language	Style and For mat	
Interpretation Strategies	30.6% (10.5)	0.2% (0.2)	1.2% (1.0)	4.6% (2.5)	26.1% (9.8)	0.7% (0.6)	*63.3% (15.5)*
Judgment Strategies	7.5% (6.1)	1.8% (1.2)	3.9% (2.0)	9.1% (3.9)	13.5% (5.9)	0.9% (0.6)	*36.7% (15.5)*
Total	*38.1% (7.9)*	*1.9% (1.3)*	*5.1% (2.7)*	*13.6% (4.6)*	*39.6% (8.0)*	*1.6% (1.0)*	*100%*

4.2.1 Interpreting the Student Scripts

Interpretation Strategies involved raters' attempts to construct an image of the script they were assessing. The raters in this study put a large amount of energy into creating an image of the scripts during the independent rating sessions. Interpretation Strategies made up 63.3% of the total number of strategies.

4.2.1.1 Reading Chunks and Treating Language Errors

To create an image of the student scripts the decision-making behavior par excellence was, not surprisingly, Read and Reread Student Script (making up 29.3% of the total number of decision-making behaviors). This strategy was accompanied by behaviors related to error treatment (making up 26.1% of the total number of behaviors), such as Edit or Correct Language (errors or unclear phrases) (M=13.4%)[42]:

> Der står ETHIC IN COMPANIES i overskriften. Og der mangler et –S i ETHICS. Det tilføjer jeg.
> Jeg starter med at læse.
> MANY COMPANIES ALL AROUND THE WORLD DECIDE, der er –S på DECIDES, jeg stryger –S'et.
> TO TRANSFER THEIR FIRM OR A PART OF THE FIRM TO ASIA OR OTHER PLACES WHERE, der er den igen, den stavefejl. Jeg tilføjer et H.
> IT IS MUCH CHEAPER TO MAKE THEIR PRODUCTS.
> Ehrm.
> AS CONSUM, ja, der er –S på CONSUMERS, vi er mere end een, vi er mange forbrugere. AS CONSUMERS, -S på
> WE ARE HAPPY TO GET vi skal også have mere end eet produkt, PRODUCTS,
> (Gitte, script 9)
>
> *Translation:*
> *It says ETHIC IN COMPANIES in the title. And an –S is missing in ETHICS. I'll add that.*
> *I'll start reading.*
> *MANY COMPANIES ALL AROUND THE WORLD DECIDE, there is an –S in DECIDES, I'll delete the –S.*
> *TO TRANSFER THEIR FIRM OR A PART OF THE FIRM TO ASIA OR OTHER PLACES WHERE, der er den igen, den stavefejl. Jeg tilføjer et H.*
> *IT IS MUCH CHEAPER TO MAKE THEIR PRODUCTS.*
> *Ehrm.*
> *AS CONSUM, yes, there is an –S in CONSUMERS, we are more than one, we are many consumers. AS CONSUMERS, add an –S*

[42] Capitalization indicates reading from the student script; underlined capitalization indicates the raters' corrected version of the student script. For more transcription conventions see Appendix E).

WE ARE HAPPY TO GET we also have to have more than one product, PRODUCTS,
(Gitte, script 9)

Classify Language Errors (*M*=7.9%),

Lad mig se på den
CONCERNING ETHICS WITHIN PRODUCTION OF ARTICLES AND SERVICES IN THE THIRD WORLD COUNTRIES, IT IS VERY IMPORTANT TO ALWAYS KEEP THE HUMAN RIGHTS IN MIND. WE HAVE OFTEN HEARD OF LARGER
Har vi så en falsk komparativ her.
COMPANIES THAT ARE EXPLOIDING
En stavefejl
...
ASIA OR EASTERN EUROPE, IT IS EXTREMELY IMPORTANT THAT YOU REMEMBER TO KEEP THE HUMAN RIGHT
Så lige en pluralis
(Torben, script 2)

Translation:
Let me have a look at it.
CONCERNING ETHICS WITHIN PRODUCTION OF ARTICLES AND SERVICES IN THE THIRD WORLD COUNTRIES, IT IS VERY IMPORTANT TO ALWAYS KEEP THE HUMAN RIGHTS IN MIND. WE HAVE OFTEN HEARD OF LARGER
We have here a false comparative.
COMPANIES THAT ARE EXPLOIDING
A spelling error
...
ASIA OR EASTERN EUROPE, IT IS EXTREMELY IMPORTANT THAT YOU REMEMBER TO KEEP THE HUMAN RIGHT
And then the plural
(Torben, script 2)

or simply Identify Errors (*M*=4.8%),

AND WHICH IS NOT.
Igen en sprogfejl
(Susanne, script 6)

Translation:
AND WHICH IS NOT.
Another language error
(Susanne, script 6)

The prevalent tendency toward error treatment is epitomized in the following extract, articulated after having read through a student script:

> Det er sådan set de fejl jeg har fundet
> (Pernille, script 5)
>
> *Translation:*
> *These are the errors I have found*
> (Pernille, script 5)

4.2.1.2 Interpreting Content and Use of Source Materials

Although reading the student script followed by error treatment or editing proved to be a typical micro sequence of behaviors, a language focus was not the only behavior to assist the reading in interpreting the scripts. Some attention was put on the content of the student scripts as well (4.6% of the decision making behaviors was devoted to content/use of source materials). The attention to the content of the script took the form of summarizing or discerning ideas (Discern or Summarize Ideas: M=1.8%):

> SO IF THE COMPANIES WANT TO BE IN LINE WITH HUMAN RIGHTS, WHAT CAN THEY DO TO PREVENT THEIR FOREIGN CONTRACT COMPANIES FROM CONTINUING EXPLOITATION? Så han siger, at hvis de gerne vil være på linie med human rights, hvad kan de så gøre for at forhindre deres udenlandske partnere i at fortsætte udnyttelsen.
> (Helle, script 12)
>
> *Translation:*
> *SO IF THE COMPANIES WANT TO BE IN LINE WITH HUMAN RIGHTS, WHAT CAN THEY DO TO PREVENT THEIR FOREIGN CONTRACT COMPANIES FROM CONTINUING EXPLOITATION? So he says that if they want to be respect human rights, what can they do to prevent their foreign contractors from continuing the exploitation.*
> *(Helle, script 12)*

When the reading process was not smooth, attempts were made to interpret ambiguous phrases (Identify or Interpret Ambiguous or Unclear Phrases: M=2.8%):

> TO THOSE COMPANIES
> Ja, igen er jeg simpelthen i tvivl om hvad meningen er. Jeg tror det er noget med det her med at de har nogle der tjekker at forholdene er i orden.
> (Thea, script 9)

Translation:
TO THOSE COMPANIES
Yes, once again I do not know that the meaning is. I think it is something about them having people that make sure that the conditions are ok.
(Thea, script 9)

4.2.1.3 Interpreting Organizational Structure and Style/Format

Although not as prevalent as the focus on language and content of the student scripts some attention was also paid to organization (Discern or Scan Organizational Structure: M=1.2%)

A LOT OF COMPANIES HAVE MOVED THEIR PRODUCTION TO OTHER COUNTRIES. THEY DO NOT THINK ABOUT THE BAD PAYMENT, CHILD LABOUR AND HUMAN RIGHTS. AND THAT IS A HUGE PROBLEM. ALL OVER THE WORLD.
Det var så appetizeren, der skulle, der skulle få os til at læse videre.
Så kommer selve det der skulle være artiklen.
(Lone, script 6)

Translation:
A LOT OF COMPANIES HAVE MOVED THEIR PRODUCTION TO OTHER COUNTRIES. THEY DO NOT THINK ABOUT THE BAD PAYMENT, CHILD LABOUR AND HUMAN RIGHTS. AND THAT IS A HUGE PROBLEM. ALL OVER THE WORLD.
This was the appetizer meant for us to be stimulated to go on reading.
Then here comes what is supposed to be the article.
(Lone, script 6)

and to the style, register, or genre (Discern Style, Register or Genre: M=0.7%):

ETHICS HAS ALWAYS BEEN A DIFFICULT THING TO DESCRIBE AND ESPECIALLY BUSINESS ETHICS. WHAT IS ETHICAL CORRECT AND WHAT IS NOT?
Så igen er vi i indledningen ind over et retorisk spørgsmål.
(Torben, script 5)

Translation:
ETHICS HAS ALWAYS BEEN A DIFFICULT THING TO DESCRIBE AND ESPECIALLY BUSINESS ETHICS. WHAT IS ETHICAL CORRECT AND WHAT IS NOT?
Again, a rhetorical question in the introduction.
(Torben, script 5)

4.2.1.4 Scanning Length

To further assist in creating an image of the scripts to be assessed, the raters also scanned the student script for length (Scan Script for Length: *M*=0.2%):

> ETHICS HAS ALWAYS,
> Nej, lad os nu se, jeg kigger lige. Hvor lang er den, halvanden side.
> BUSINESS ETHICS IN THE THIRD WORLD, ja.
> (Nina, script 5)
>
> *Translation:*
> *ETHICS HAS ALWAYS,*
> *No, let me see, I am looking. How long is it, a page and a half.*
> *BUSINESS ETHICS IN THE THIRD WORLD, yes.*
> *(Nina, script 5)*

An average 0.2% of the total decision-making behaviors devoted to scanning the script for length may seem trivial. However, considering that it doubtlessly only takes about one (or few) decision-making behavior(s) to get an image of the length of a script, and likely more than one decision-making behavior for the content and the language of a script, consideration of composition length seemed to be part of raters' attempts to create an image of a script.

4.2.1.5 Interpreting Context and Monitoring Self

When not focusing directly on the textual features of the student scripts, the raters sometimes sought to envision the student's personal situation (Envision Personal Situation of the Student: *M*=0.8%). In this regard, the raters sometimes considered time restrictions:

> Der mangler flere ord. Eleven er lidt presset for tid.
> (Astrid, script 1)
>
> *Translation:*
> *Several words are missing. The student is under time pressure.*
> *(Astrid, script 1)*

They also wondered where the students got their ideas from:

> Ja, jeg sidder lige og kigger, tænker på, her når det bliver skrevet om Nike i de sidste 8 linier af andet afsnit, at det lyder meget som noget der har været brugt før. Man kunne godt forestille sig at eleven har skrevet en opgave om det her før og sidder og genbruger fra tidligere opgaver.

(Malene, script 7)

Translation:
Yes, I am looking and thinking that what is being written about Nike in the last 8 lines of the second paragraph, that it sounds as if it has been used before. It could be that the student has written an essay about this before and then is rewriting this here.
(Malene, script 7)

Or they considered how the student may have interacted with the computer:

NOT LEAST BECAUSE OF THE INFLUENCE FROM AMNESTY INTERNATIONAL BECAUSE COMPANIES
Se der er et stort B alene fordi det beder computeren om efter et punktum. Så gør den det automatisk.
(Lone, script 12)

Translation:
NOT LEAST BECAUSE OF THE INFLUENCE FROM AMNESTY INTERNATIONAL BECAUSE COMPANIES
See that is a capital B, merely because the computer does it after a period. It is automatic.
(Lone, script 12)

Related to envisioning the personal situation of the student, raters considered the task or exam requirements the students were subjected to (Consider Task or Exam Requirements: M=0.2%):

Måske er den opgaveformulering ikke bred nok. Der står immervæk også kun en linie og et ord sammenlignet med hvis man havde valgt A-opgaven. Det er altså lidt noget andet. Der er meget mere. Der får man structuren foræret og mange flere eksempler. Man får ingenting her Det kan de åbenbart ikke magte.
(Lone, script 5)

Translation:
Perhaps the essay prompt is not broad enough. It is, in fact, only one line and one word compared to the A-prompt. So it is a bit different. There is a lot more. They are given the structure and many more examples. Here they get nothing. Apparently, that is difficult for them.
(Lone, script 5)

Some very conscientious raters felt the need to make sure that they were creating the right image of the scripts, taking effort to consult other sources in cases where they were unsure of their own capabilities in English (Consider Own Perception of Correct English (e.g., consult a dictionary): M= 0.2%):

Ja, så står der igen ETHICS HAS. Jeg tror lige for en sikkerheds skyld, at jeg vil slå op om det ikke er rigtigt at ETHICS skal være i forbindelse med flertal, så jeg ikke retter noget det er, som alligevel er rigtigt. Jeg slå op I en engelsk-engelsk ordbog, McMilllan, og skal finde ETHICS, og det har vi her "ethics". Der er flertal "principles that are used to decide" ehrm "what is right and what is wrong". Ja, så det må skulle forbindes med flertal. Jeg lægger ordbogen væk".
(Malene, script 5)

Translation:
Yes, again it says ETHICS HAS. I think I will look it up to make sure that ETHICS must be combined with a plural [form of the verb], so that I am not correcting something that is, in fact, correct. I am looking it up in a monolingual dictionary, McMillan, finding ETHICS, here we have it "ethics". It is plural "principles that are used to decide" ehrm "what is right and what is wrong". Yes, it must be combined with a plural [verb]. I am putting the dictionary away.
(Malene, script 5)

4.2.1.6 Summary of Raters' Interpretation Strategies in Independent Ratings

In sum, to create an image of the student scripts, the raters typically read chunks of the scripts followed by behaviors related to error treatment (editing or correcting language errors, classifying errors, or merely identifying errors). Attention was also drawn to the content of the scripts. More specifically, the raters would attempt to discern or summarize the student's ideas, sometimes making an effort to interpret ambiguous phrases. Although not particularly prevalent in the raters' protocols, organizational structure, style/format and length also seemed to catch the raters' attention. To assist the textual focus in interpreting the student scripts the raters looked to contextual factors as well as their own self-monitoring behaviors: In particular, the personal situation of the student was considered, including time restrictions, essay prompt, and use of computer. But the raters also at times made an effort to monitor their own strategies by making sure they did not regard correct language as incorrect.

4.2.2 Judging the Student Scripts

To judge the image they had created of the student scripts, the raters employed a wide range of judgment strategies. As will be apparent later on (in reporting the sequence of behaviors), the raters often interspersed these judgment strategies with their interpretation strategies.

4.2.2.1 Judging Context and Self

In judging the scripts the raters, not surprisingly, articulated a score. The behavior, Articulate Score (M=3.1%) was often accompanied by a general impression (Articulate General Impression: M=1.8%):

> Jeg vil alligevel vælge at holde opgaven på en 8'er, som det kommer nærmest. En stor 8'er vil jeg holde opgaven til. Det er pænt og ordentligt, men der er ikke det løft der gør det til et essay på et højere niveau.
> (Hans, script 12)
>
> *Translation:*
> *I will still give the essay an 8, as that is what comes closest. I will give it a high 8. It is nice and neat, but it doesn't have this extra edge that pushes it to a higher level.*
> *(Hans, script 12)*

The raters supported their general impressions with direct examples from the student scripts, making sure their judgments could be justified by concrete examples in the communal rating session to come (Exemplify Directly From Student Script: M=1.4%):

> Og ehrm flere af de grammatiske fejl er deciderede meningsforstyrrende. Altså PRODUCTS ARE MADE BY HUMAN RIGHTS, for eksempel.
> (Jesper, script 14)
>
> *Translation:*
> *And ehrm several of these grammatical errors, in facts, disturbs the meaning. Well, PRODUCTS ARE MADE BY HUMAN RIGHTS, to take an example.*
> *(Jesper, script 14)*

The following statement by Tina indicates distinctly that the raters were preparing themselves for justifying their communal assessments by making sure they could provide concrete documentation:

> Gælder om at tage nogle notater til den fælles evaluering.
> (Tina, script 1)
>
> *Translation:*
> *It is all about taking notes in preparation for the communal rating session.*
> *(Tina, script 1)*

I had expected the raters to directly consult the scale descriptors when in doubt about which final score to assign. However, they never directly consulted the descriptors. Whether the raters did not make scale-related comments because they had internalized the scale or whether they chose to deliberately ignore it because they did not find it useful is hard to tell[43]. Nevertheless, when insecure about which exact score to assign (or other insecurities for that matter) they resorted to other sources of assistance. They would sometimes compare with other scripts they were rating (Compare Student script: M=0.7%):

> Så jeg er nede ved den lave ende. Jeg kan ikke rigtig bestemme mig til om det skal være et ... Jeg synes den skal bestås, ikke? Det er, vi taler 8. Spørgsmålet er så om vi skal op og tale 7. Jeg mener faktisk der var. Var der ikke en anden jeg gav 7 også? Jeg tager lige de første jeg rettede. Jeg tager lige dem frem.
> Jo, den aller, allerførste gav jeg 7.
> Ehrm, og det var den der INCOME RAISE – GOOD SENSE GONE. Den kan jo godt minde lidt om den her, nævner også McDonald's og så videre. Så ehrm, den var jeg flink ved. Det må jeg også hellere være ved denne her. Så jeg giver et 7-tal.
> (Torben, script 15)
>
> *Translation:*
> *Now I am in the lower end of the scale. I cannot really make up my mind whether it should be a... I think it is a pass, right? It is, we are talking 8.*
> *The question is then whether we should go up, around a 7. In fact, I think it was. Didn't I give a 7 to another one too? I am just going to look at the first ones I graded. I am taking them out.*
> *Yes, the very, the very first one I gave a 7.*
> *Ehrm, and it was this INCOME RAISE – GOOD SENSE GONE. It bears some resemblance to this one, also mentioning McDonald's and so on. So ehrm, I was lenient with that one. I'd better be lenient with this one, too. So I'll give it a 7.*
> *(Torben, script 15)*

Or they would leave the score decision to the communal rating session:

[43] The retrospective questionnaire, however, reveals that raters find the communal rating sessions more useful than a rubric. As one rater put it:

> Det [fællesbedoemmelser] er vigtigt, fordi censorer skal helst have så meget fælles grundlag som muligt, og der er mange ting, vi ikke kan stille regler for. Man kan ikke bare sådan give et tal. Derfor er samtale vigtigt.
> (Pernille, retrospective questionnaire)
> *Translation:*
> *It is important because raters must judge by the same standards. There are so many things you cannot make rules for. You cannot just assign a score like that. That is why conversation is important.*
> *(Pernille, retrospective questionnaire)*

> Jahh, den er på 9 til 10, og så vil jeg lade medcensor afgøre hvor vi skal hen. Det kan også godt være vi skal i en hel anden retning. Det må vi vente til.
> (Jesper, script 13)
>
> *Translation:*
> *Yeah, it is a 9 to 10, and then I will let my co-rater decide which way to go. It could also be that we should go in a totally different direction. We have to wait and see.*
> *(Jesper, script 13)*

The communal rating sessions to come were seen not just as a score resolution method, but also as a forum in which other insecurities could be resolved:

> Det der kan være lidt et problem det er at ehrm at grundteksterne ikke bliver brugt så forfærdeligt meget, og slet ikke den den danske grundtekst. Og det har jeg lige skrevet et spørgsmålstegn til på min blok, og det er så noget noget jeg må snakke med [Thea] om hvor meget vi vil vi vil straffe det. Jeg bliver nødt til at gå videre til den næste opgave.
> (Malene, script 5)
>
> *Translation:*
> *It may be a bit of a problem that ehrm that the source texts are not used much, and not the Danish source text at all. And I have noted that with a question mark, and that that I need to talk to [Thea] about how much we should punish the student for this. I have to go on to the next essay.*
> *(Malene, script 5)*

They were also perceived to be an opportunity to align harshness or leniency:

> Ehrm den er rimelig avanceret i sin sætningsstruktur, og hvis der havde været skrevet to sider, ville jeg slet, slet ikke have været i tvivl. Hvis der havde værte skrevet to fulde sider, ville jeg ikke være i tvivl om at det var en klokkeklar 11. Men pga det lidt manglende omfang vil jeg give den 10, og jeg mener der skal ske et fradrag, så det ender med 10. Det kan godt være det er urimeligt, men så må medcensor korrigere. Men jeg mener vi skal holde fast i at de skal skrive lidt mere end der er skrevet her.
> (Jesper, script 12)
>
> *Translation:*
> *Ehrm, the sentence structure in this is rather advanced, and had there been two pages, I wouldn't have been in doubt at all. Had there been two full pages, I wouldn't have doubted that this was a clear 11. But because of the rather small amount of text I'll give it a 10, and I think that points should be taken off, so I end up with a 10. That might seem unreasonable, but then my co-rater must correct me. But I think we must maintain a requirement that they have to write more than what is written here.*
> *(Jesper, script 12)*

In fact, to signal that their independent scores were open for negotiation all raters articulated flexible scores, such as "a high 8", "a low 7" or "a 6 or a 7". Of the 300 scores (20 raters rating 15 scripts) assigned in the independent rating sessions, 128 scores were flexible in this manner. See Appendix H for all scores assigned.

The raters were, indeed, often quite explicit about their assessment strategies (Define, Revise, or Suggest Assessment Strategies: M=0.4):

> Lad os lige se hvad der står i opgaven. Den er halvanden side. Egentlig så plejer jeg at sige at tænke at under halvanden side, så er det automatisk et 6-tal fra starten, men det kommer også an på hvad de får besked af deres lærer på at skrive, hvor meget der er ok, så det kan jeg ikke klantre eleven for.
> (Nina, script 3)
>
> *Translation:*
> *Let us see what it says in the essay. It is a page and a half. As a matter of fact, I usually say think that anything less than a page and a half is automatically a 6 from the beginning, but it also depends on what the teacher tells them to write, how much is ok, so I cannot blame the student for that.*
> *(Nina, script 3)*

4.2.2.2 Judging Language

As was the case when interpreting the scripts, the raters paid special attention to features related to language when judging their constructed images of the scripts. As can be seen from Table 4.2, 13.5% of the decision-making strategies were devoted to judging the language in the scripts. Often the raters judged the student's overall language competence (Assess or Justify Language Overall: M=2.6%):

> Ja, jeg sidder og tænker på at den her opgave har også et rigtigt fint sprog.
> Og jeg læser videre
> (Malene, script 6)
>
> *Translation:*
> *Yeah, I am thinking that this essay also has a really nice language.*
> *And I am reading on*
> *(Malene, script 6)*

More specifically the raters assessed the students' fluency or comprehensibility level (Assess or Justify Fluency or Comprehensibility: M=3.4%). This was sometimes done in positive terms:

> THAT HAS PUT GREAT EMPHASIS ON GIVING THE WORKERS THE BEST CONDITIONS. THEY MAKE SURE THAT THE WORKERS ARE TREATED IN THE SAME WAY AS THE ONES WHO WORK AT THE FACTORIES BASED IN AMERICA.
> Meget, meget fin fluency i det der
> (Hans, script 2)
>
> *Translation:*
> *THAT HAS PUT GREAT EMPHASIS ON GIVING THE WORKERS THE BEST CONDITIONS. THEY MAKE SURE THAT THE WORKERS ARE TREATED IN THE SAME WAY AS THE ONES WHO WORK AT THE FACTORIES BASED IN AMERICA.*
> *Very, very nice fluency in this one*
> *(Hans, script 2)*

As well as in negative terms:

> IN DENMARK, CHILD LABOUR IS NOT A PROBLEM, THE DANISH PEOPLE ARE VERY MUCH AGAINST IT, AND THEY DID RATHER PAY THE CHILDREN RIGHT
> Uha, formuleringsproblemer. Alvorligt.
> (Astrid, script 8)
>
> *Translation:*
> *IN DENMARK, CHILD LABOUR IS NOT A PROBLEM, THE DANISH PEOPLE ARE VERY MUCH AGAINST IT, AND THEY DID RATHER PAY THE CHILDREN RIGHT*
> *Oh, problems making himself understood. Serious problems.*
> *(Astrid, script 8)*

Some raters stated that comprehensibility was a very important issue:

> Altså, meningen, jeg forstår hvad han siger, men det er totalt dårligt skrevet. Så vi er stadigvæk på et 6-tal.
> (Nina, script 5)
>
> *Translation:*
> *Well, the meaning, I understand what he is saying, but it is poorly written. So we are still on a 6.*
> *(Nina, script 5)*

One rater, in particular, was particularly annoyed with students' incomprehensible language:

> SO MUCH MONEY ON THIS PROJECT WAS FAILURE, THEN HOW ARE WE AS HUMANS LOOKING AT THIS

> Altså, det sprog er meget dårligt, og jeg er ikke sikker på hvad det skal betyde. Og som jeg sagde før, vil jeg altså ikke bruge mere tid til at tænke på hvad kunne det evt. have ment for jeg synes altså at det er for dårligt.
> (Pernille, script 14)
>
> *Translation:*
> *SO MUCH MONEY ON THIS PROJECT WAS FAILURE, THEN HOW ARE WE AS HUMANS LOOKING AT THIS*
> *Well, this language is very poor, and I don't even know what it is supposed to mean. And as I said before, I really don't want to spend more time figuring out what it might mean, because it really is too bad.*
> *(Pernille, script 14)*

Attention was also paid to syntax or morphology (Assess or Justify Syntax or Morphology: M=2.5%):

> ONLY FOR A TEMPORARY PERIOD.
> OK, det virker som om der er styr på EASILY, TEMPORARY, osv.
> Noget med adjektiver og adverbier Laver et '+' i margen.
> (Louise, script 2)
>
> *Translation:*
> *ONLY FOR A TEMPORARY PERIOD.*
> *Okay, it seems as if he masters EASILY, TEMPORARY, etc.*
> *Something with adjectives and adverbs. Writing a '+' in the margin*
> *(Louise, script 2)*

Often the attention to syntax or morphology was judged on the basis of an assessment of the errors encountered in the script. In particular, the frequencies of errors (Assess or Justify Frequency of Errors: M=2.1%) were evaluated and seemed to play a particularly important role in judging the script:

> PRODUCTION OF ARTICLES AND SERVICES IN THE THRID WORLD
> Det er en god opgave. Der er ikke ret mange fejl.
> (Pernille, script 7)
>
> *Translation:*
> *PRODUCTION OF ARTICLES AND SERVICES IN THE THRID WORLD*
> *It is a good essay. There are few errors.*
> *(Pernille, script 7)*

The gravity of errors also seemed to be an issue: (Assess or Justify Gravity of Errors: M=0.8%), often combined with an assessment of the frequency of errors:

Nå, det var den. Hvad siger vi her.
Jeg kigger lige på de sproglige fejl, hvad det er for nogle streger jeg har sat her. Dem er der en del af, synes jeg. Og nogle også af de alvorlige fejl, som jeg har sat to streger under.
Ehrm. Ehrm. Uden at det jo er helt umuligt, så sprogligt ligger den vel på omkring 7, vil jeg sige. Og det er nok den den lander på.
(Thea, script 1)

Translation:
Well, that was this one. What do we say here?
I am looking at the language related errors, what type of errors have I noticed. There is a lot of them, I think. And some of them also severe. I have underlined those twice.
Ehrm, ehrm. It isn't impossible, so the language is about a 7, I would say. And this is what I will end up with.
(Thea, script 1)

Lexis was also taken into consideration (Assess or Justify Lexis: M=1.4%). The raters were sometimes taken by the use of a single word or a phrase, as in:

Pænt sprog OUTSOURCED. Det er et fint ord at bruge, et relativt nyt begreb, og han burger det korrekt her. Det tyder på at hun har antennerne slået ud. Det er udmærket
(Astrid, script 12)

Translation:
Nice language OUTSOURCED. This is an advanced word to use, a relatively new term, and he is using it correctly here. She appears to notice things. That is excellent.
(Astrid, script 12)

And/or by the overall command of vocabulary:

Rent sprogligt er det også et primitivt sprog med et meget begrænset ordforråd og utrolig mange fejl. Så 5, i bedste tilfælde 6. Det er nok min umiddelbare holdning.
(Susanne, script 1)

Translation:
Looking at the language: it is a primitive language with a very limited vocabulary and an excessive amount of errors. So 5, a 6 at best. That is probably my immediate assessment.
(Susanne, script 1)

4.2.2.3 Judging Content and Use of Source Materials

Although language-related features of the student scripts took up a relatively large portion of the raters' attention, raters attended also to the students' ideas and their use of the source materials in the exam packet given to them. The raters' consideration of ideas and use of the source materials made up 9.1% of the raters' decision-making strategies.

Attention to content-related features of the scripts was sometimes paid in the form of an overall assessment of the content or ideas (Assess or Justify Content/Ideas Overall: *M*=1.2%):

> Jeg tænker lidt indtil videre at det er en opgave der lyder sådan indholdsmæssigt som om at den har fat i noget af det rigtigte.
> (Jens, script 1)
>
> *Translation:*
> *So far I am thinking that looking at the content, this is an essay where the student knows what it is all about.*
> *(Jens, script 1)*

Or sometimes an assessment was made of how well the student had used the accompanying source materials (Assess or Justify Use and Understanding of Source Material: *M*=1.7%). The focus could be on how well the student appeared to have understood the materials:

> DESTROYED, AND THE WOULDN'T USE THE COMPANY ANY MORE.
> Meget primitive fremstilling af det der står i teksen.
> REEBOK HAVE TRIED TO DO MANY THINGS TO BE LEGAL.
> Ja, det er heller ikke snedigt fremstillet.
> (Astrid, script 9)
>
> *Translation:*
> *DESTROYED, AND THE WOULDN'T USE THE COMPANY ANY MORE.*
> *Very primitive account of what the source text says.*
> *REEBOK HAVE TRIED TO DO MANY THINGS TO BE LEGAL.*
> *Yes, that is not sharp either*
> *(Astrid, script 9)*

Or raters focused on how the students managed to incorporate useful information from the source materials into their scripts:

> Lad os se, hvad var det der stod her? Det er det her med, altså meget referende af teksten egentlig, ikke? Fremlægger problemet, det er det der med at der er børnearbejde. Og så hører vi om at en en ung fyr.
> Sidder lige og læser igen,
> Ehrm der arbejder 10 til 14 timer om dagen og ikke fik nogen betaling. Så kommer der den der fuldstændige ligegyldige der.
> (Thea, script 9)
>
> *Translation:*
> *Let us see, what did it say her? This thing about, really pretty much a summary of the text, isn't it? Gives an account of the problem, the thing about child labor. And then we hear about a young guy.*
> *Reading it again.*
> *Ehrm, who is working 10 to 14 hours a day and didn't get paid. Then comes something completely irrelevant there.*
> *(Thea, script 9)*

Some raters were satisfied even with just a mention of the source texts.

> THE MANAGER OF THE FACTORY HAVE TO INFORM THEM ABOUT THE REASON AND IMPORTANCE WHY THEY HAVE TO BE A MEMBER OF A TRADE UNION.
> Ja, godt. Der var noget ind fra artiken. Jeg skriver 'godt' i margen.
> (Louise, script 1)
>
> *Translation:*
> *THE MANAGER OF THE FACTORY HAVE TO INFORM THEM ABOUT THE REASON AND IMPORTANCE WHY THEY HAVE TO BE A MEMBER OF A TRADE UNION.*
> *Yes, good. There was something in here from the article. I am writing 'good' in the margin.*
> *(Louise, script 1)*

Besides use of the source materials, some raters assessed the students' sense of independent thinking,

> Altså man må jo så sige at hvis man sammenligner med med opgave 10. Den er jo så på en eller anden made mere selvstændig den her, for der er jo ikke noget med citater eller den slags.
> Ja, den er svær at bedømme.
> (Jette, script 11)
>
> *Translation:*
> *Well, I must say that if I compare it with essay 10, then it is somehow more independent this one, because there are no quotations or the like.*
> *Yes, it is difficult to rate.*
> *(Jette, script 11)*

and the sense of maturity they expressed (Assess or Justify Maturity or Independence: *M*=0.9%):

> NO ONE FEELS COMEFORBABLE TALKING ABOUT THIS SUBJECT, BECAUSE IT INVOLVES LITTLE CHIDLREN WHO SUFFER, AND THE FACT THAT NO PEOPLE ARE ABLE TO DO ANTHING ABOUT IT, MAKES THEM (US) NOT WANT TO TALK ABOUT THIS EMBARASSING SUBJECT, THAT IT HAS BECOME
> Nej, det er rigtig nok ubehageligt, men der er netop en masse tale om det for tiden. Ehrm, det er sådan at han prøver at gøre sig selv *. Men det er også igen lidt langt fra modenhed.
> (Helle, script 3)
>
> *Translation:*
> *NO ONE FEELS COMEFORBABLE TALKING ABOUT THIS SUBJECT, BECAUSE IT INVOLVES LITTLE CHIDLREN WHO SUFFER, AND THE FACT THAT NO PEOPLE ARE ABLE TO DO ANTHING ABOUT IT, MAKES THEM (US) NOT WANT TO TALK ABOUT THIS EMBARASSING SUBJECT, THAT IT HAS BECOME*
> *No, this is really uncomfortable, but there is a lot of talk about it nowadays. Ehrm, he also tries to *. But it is also far from mature.*
> *(Helle, script 3)*

Topic development seemed to be an issue too:

> Det er en god ide taktisk ide, synes jeg, at stille spørgsmål, hvis man gerne vil have ordet selv. * kan man få svaret, og det er det eleven lægger op til. Men eleven kommer allerede i tredie linie med et svar, som jeg synes er en lille smule overfladisk, fordi de ikke vil have dårlig samvittighed. Jeg ved ikke helt om om man allerede så hurtigt kan besvare det. Her sætter jeg en lille bølgestreg ehrm. Jeg er ikke ehrm tilfreds med argumentet.
> (Astrid, script 3)
>
> *Translation:*
> *It is a good idea tactically, I think, to ask questions if you want the word * you can get the answer, and that is what the student is trying to do. But the student comes up with an answer already in the third line, which I think is a bit superficial, because they don't want a bad conscience. I don't really know if one could answer this so soon. I will underline this ehrm. I am not satisfied with the argument.*
> *(Astrid, script 3)*

So did the students' reasoning or logic (Assess or Justify Reasoning, Logic, or Topic Development: *M*=1.6%):

> REEBOK DID DO SOMETHING
> Jeg synes jo netop her at vedkommende modsiger sig selv; han har lige skrevet om at man kan ikke gøre noget, og det hjælper heller ikke noget, og så står der alligevel her at Reebok har gjort noget. Og så står der i næste linie, men det er ligegyldigt
> IT MAKES NO DIFFERENCE
> Nå

(Pernille, script 3)

Translation:
REEBOK DID DO SOMETHING
Right here I think the test taker contradicts himself; he has written something about people not being able to do anything, and apparently it doesn't help. And then it says in the next line, but that doesn't matter
IT MAKES NO DIFFERENCE
Well
(Pernille, script 3)

and the composition's relevance to the assigned essay topic (Assess or Justify Task/Topic Relevance or Completion: M=2.0%)

Det er jo noget væk fra emnet. Men jeg kan da godt forstå vedkommende har skrevet det, for det er jo netop det her med forskellige religioner kan se forskelligt på tingene. Så derfor kan man sige, at 9/11 kommer jo ind rigtigt nok som et eksempel her. Men vedkommende kan jo også godt se at det er jo ikke lige det der skal diskuteres, og kommer så altså også tilbage til til emnet i fjerde afsnit, fordi han/hun skriver at det er jo bare et eksempel, og at man skal om child workers. Men det skulle man jo egentlig heller ikke. Der er jo en masse andre ting involveret.
(Pernille, script 3)

Translation:
It is pretty far away from the topic. But I do understand why the test taker mentioned it, because it is exactly the thing about different religions that they view things from different perspectives. So that's why we could say that 9/11 fits well in here as an example. But the test taker should know that it doesn't exactly fit into the discussion right here, and he does return to to the topic in the fourth paragraph, because he/she writes that it is just an example, and that we should about child workers. But, in fact, we shouldn't. Many other things are involved.
(Pernille, script 3)

Reactions also appeared to the student's viewpoints. The raters sometimes expressed disagreement with the student's opinion or his/her statements (Assess or Justify Correctness of or Disagreement with Content: M=1.5%):

CHILDREN WORK IS SOMETHING YOU DO NOT SAY OUT LOUD OR AIR YOUR OPINION ABOUT.
Jeg mener da ikke det er rigtigt. Det synes jeg da netop der tales meget om idag. Og der er mange firmaer der er begyndt at gøre noget ved det. Fordi så kan de jo stadigvæk blive ved med at tjene deres gode penge på billig, underkuet arbejdskraft.
(Pernille, script 3)

Translation:

CHILDREN WORK IS SOMETHING YOU DO NOT SAY OUT LOUD OR AIR YOUR OPINION ABOUT.
I don't think that is true. In fact, I think that people often do talk about it today. And many companies have started doing something about it. Because then they can still keep on earning their good money using cheap, suppressed labor.
(Pernille, script 3)

4.2.2.4 Judging Organizational Structure

Often related to Content (especially topic development), Organizational Structure was taken into consideration when judging the scripts (*M*=3.9% of the total amount of decision-making behaviors were devoted to judging organizational structure). This was sometimes done by assessing the overall organizational structure (Assess or Justify Organizational Structure Overall: 0.7%). On a more specific level, the raters would assess the title[44] given to the composition (Assess or Justify Title: *M*=1.2%),

Overskriften er heller ikke en overskrift der egentlig giver god information til det man nu kan forvente at lease.
(Susanne, script 1)

Translation:
Also, the title is not a title that really gives us valuable information about what we can expect to read.
(Susanne, script 1)

Or the introduction,

Indledningen der er et billedligt spørgsmål.
HAVE YOU EVER THOUGHT ABOUT WHERE YOUR BRAND NEW SWEATSHIRT, JEANS OR SNEAKERS ARE PRODUCED AND BY WHOM? OR DO YO KNOW ANYONE WHO HAS?
Det er egentlig en ret god introduction. Og så bliver der åbnet op for hvor denne production kan finde sted
(Susanne, script 3)

Translation:
The introduction, which is a question that provides us with a picture.
HAVE YOU EVER THOUGHT ABOUT WHERE YOUR BRAND NEW SWEATSHIRT, JEANS OR SNEAKERS ARE PRODUCED AND BY WHOM? OR DO YO KNOW ANYONE WHO HAS?

[44] The students had to create an appropriate title for their essays.

It is, in fact, a rather good introduction. And then it paves the way for where this production could take place.
(Susanne, script 3)

Or the conclusion,

Knap så elegant afsluttet, selvom den sidste sætning afslutter essayet på en rimelig facon. Det er lidt specielt at vedkommende siger at
IT IS DEVASTATING THAT NOTHING CAN BE DONE TO HELP SOLVING THE PROBLEM
(Henrik, script 4)

Translation:
Not a particularly elegant conclusion, although the last sentence concludes the essay acceptably. It is a bit odd that the student says
IT IS DEVASTATING THAT NOTHING CAN BE DONE TO HELP SOLVING THE PROBLEM
(Henrik, script 14)

The behavior of Assess or Justify Introduction or Conclusion made up 1.3% of the total strategies.

Assessing coherence or cohesion was also at times a concern:

Der er brugt magne connectives, der gør at den er nem at læse og det hele hænger godt sammen. Strukturen er i orden.
(Tove, script 4)

Translation:
Many connectives are used. This makes it easy to read, and there is good coherence. The organizational structure is okay.
(Tove, script 4)

But it seemed to make up a relatively small amount of the raters' judgment comments (Assess or Justify Coherence and/or Cohesion: M=0.8%).This may not be because the raters' did not pay attention to such features in that judgments about coherence might be included in their assessment of the overall organizational structure, or of the conclusion.

4.2.2.5 Judging Style and Format

Style and genre also came to play a role, although a minor one (M=0.9% of the total amount of decision-making strategies was devoted to judging Style or Genre):

> Jeg tror jeg vil give den opgave her et 7-tal. Ikke mindst fordi vedkommende altså skriver en artikel og ikke et essay, og på den måde ikke får opbygget ehrm det som de skal som et essay, så her vil jeg faktisk give et 7-tal. Måske et sted mellem. Jeg tror jeg vil holde mit et et 7-tal. Og det var så opgave nummer 8.
> (Henrik, script 8)
>
> *Translation:*
> *I think I will give this essay a 7. Not least because the test taker writes an article and not an essay, and in this way doesn't succeed in structuring ehrm it, as they have to, as an essay, so here I would, in fact, give it a 7. Perhaps somewhere in between. I think I'm going to stick to a 7. And this was essay 8.*
> *(Henrik, script 8)*

4.2.2.6 Judging Amount of Text

Judging the Amount of Text in the student scripts made up 1.8% of the total strategies. This was an issue commented on by all raters:

> Og det vil sige at vi har altså her en opgave på elleve, femten, tyve, treogtyve linier, hvoraf de syv linier er et citat, og ehrm de tal, der bruges i citatet, de bliver udelukkende kommenteret med
> THEY ARE HORRIFYING, EVEN DISGUSTING
> Ehrm, det er det eneste kommentar der på det.
> Så opgaven er jo, er jo klart uacceptabel i omfang.
> (Hans, script 10)
>
> *Translation:*
> *And it means we are dealing with an essay consisting of eleven, fifteen, twenty, twenty-three lines, seven lines of which are a quote, and ehrm the numbers that are used in the quote, they are commented on only by*
> *THEY ARE HORRIFYING, EVEN DISGUSTING*
> *Ehrm, that is the only comment on that.*
> *So the essay is, in fact, clearly unacceptable in amount of text.*
> *(Hans, script 10)*

Sometimes the amount of language written seemed to be a decisive factor in the decision-making:

> Det er en 5-6 stykker igen. Men spørgsmålet er om den kan bestås. Altså det er lidt over en halv side. Den kan godt ske at jeg er lidt for hård, men det bliver et 5-tal igen.

(Torben, script 14)

Translation:
It is a 5 or a 6 again. But the question is whether it can pass. It is just over half a page. Perhaps I am a bit harsh, but it will be yet another 5.
(Torben, script 14)

4.2.2.7 Summary of Raters' Judgment Strategies in Independent Ratings

In judging the image the raters had created of the student scripts they used a diverse set of judgment strategies. They attended to language-related features of the scripts, both their overall impressions of the students' command of language and, more specifically, to the frequency and gravity of errors as well as fluency and comprehensibility. Content and the student's use of source materials were also judged: This involved judging how the students used and understood the source texts as well as their logic or reasoning and topic relevance. The raters also judged the students' level of maturity and decided to what extent they could agree or disagree with their viewpoints. Organizational Structure and Style/Format were also judged, although Style and Format seemed to have a minor impact on the raters' judgments. The Amount of Text was also commented on to some extent, and it seemed to be a decisive factor in the final judgment of some of the scripts.

To help form and finalize their judgments or resolve their insecurities the raters seemed to not consult the scale descriptors directly (although they might have internalized them to the extent that commenting on them was so automatic that it did not warrant commentary). Rather, they would resort to other judgment strategies such as comparing with other scripts or relying on the co-rater discussions in the communal rating sessions to come. They deliberated on their own assessment strategies and sometimes expressed insecurities with them. Besides articulating a general impression of the student scripts the raters also noted examples directly from the scripts, making sure their assessments could be validated by concrete documentation in the communal rating sessions to follow.

4.3 Sequence of Decision-Making Behaviors

My coding scheme did not presume a fixed sequence of behaviors, but it soon became apparent that the raters displayed typical sequences of behaviors. There was the micro-level sequence of reading a piece of the student script and subsequently treating language errors and/or discerning or interpreting ideas, and interspersing that with judgment comments. In addition, the raters displayed a typical rating sequence on a macro level. The raters would usually go through the following three phrases in their independent rating sessions:

Phase 1: Form initial impression;
Phase 2: Interpret script while building up judgments of it;
Phase 3: Finalize preliminary scoring.

The transitions from one phase to another were in all instances very clear, as the following excerpt illustrates:

Phase 1: Tager et overblik over stilen,
kigger på længden, virker til at være lidt for kort. Jeg anslår den til en 450
Jeg bemærker at der på side to øverst er faldet en linie ud i kopien.
Kigger på overskriften, er i orden.

Phase 2: Går i gang med afsnit 1.
Læser igennem
Bemærker linie 2: slåfejl, stavefejl, 480 ord max.
TO UNDERSTAND
...
ellers rimelig
...
ikke overvældende spændende
afsnint to
læser igennem
...
udmærket første afsnit
...
IT'S NOT EVEN THEM WHO WANT TO WORK
FAMILY WHO IS GIVING
Understreger begge steder
Kongruensfejl
Næste afsnit starter med OFF CAUSE, stavet forkert, understregning
...
THE NEXT CHILDREN STANDS,
Igen en kongruensfejl
Går til afsnit tre

Om Reebok
Læser igennem
Igen kongruensfejl: THE EMPLOYEES AT THE FACTORY STANDS
THE EMPLOYEES STANDS
THE FACTORY NEEDS SOME ADJUSTMENT
Lidt uklart
F.EX.
Danisme
THE CHANCE FOR THE DEALS TO WENT THOUGH
Helt grusomt
Vi er under middel
Jeg vil skyde på en 6'er.
Går til afsnit fire
CHILD WORK THOUGH NEWS
Skal være THROUGH NEWS og så videre.
...
THE NEXT DAY YOU ARE HEARING ARE HEARING
Sjusk
Går over til næste side, starter afsnit 5
BEFORE WE ARE MAKING ANY PREMATURE CONCLUSIONS
...
JUST STOPPING
Der mangler noget der
GO AND LEAVING THEM.
Problemer med –ing-formen
Uklart afsnit
MANY CHILD AND THEIR FAMILY WILL STARVE AND POSSIBLE
Måske POSSIBLY
...
afsnit 6, det sidste afsnit
INSTEAD OF STOP, skulle være STOPPING, sætter streg under.
Og sidste linie kongruens:
THAT HAVE TO BE TURNED.

Phase 3:Jeg ser tilbage og tager et overblik over hele besvarelsen
Overvejer 6, måske 7, men 7 er nok for meget, hvis jeg kigger på de alvorlige grammatiske fejl, dumme stavefejl.
Jeg vil vurdere opgaven til en 6, pil op.
Den holder rimelig, men ikke til middel
Sprogligt for mange fejl til at kunne holde en 7'er, 6 pil op.
Færdig med opgave 1.
(Ken, script1)

Translation:

Phase 1:Taking an overview of the essay,
Looking at the length, appears to be a bit too short. Probably 450.
I notice that a line is missing on top of page two in the copy.
Looking at the title, is okay.

Phase 2:Starting paragraph one.

Reading it
Noticing line two, typos, spelling mistakes, 480 words at most.
TO UNDERSTAND
...
Otherwise reasonable
...
Not particularly exciting
Paragraph two
Reading through
...
Excellent first paragraph
...
IT'S NOT EVEN THEM WHO WANT TO WORK
FAMILY WHO IS GIVING
Underlining both places
Subject-verb agreement error
Next paragrapah starts OFF CAUSE, spelled wrong, underlining
...
THE NEXT CHILDREN STANDS,
Again a subject-verb agreement error
On to paragraph three
Again a subject-verb agreement error
On to paragraph three
About Reebok
Reading through
Again a subject-verb agreement error: THE EMPLOYEES AT THE FACTORY STANDS
THE EMPLOYEES STANDS
THE FACTORY NEEDS SOME ADJUSTMENT
A little unclear
F.EX.
Danism
THE CHANCE FOR THE DEALS TO WENT THOUGH
Completely terrible
We are below average
I would guess a 6.
On to paragraph four
CHILD WORK THOUGH NEWS
Should be THROUGH NEWS and so on.
...
THE NEXT DAY YOU ARE HEARING ARE HEARING
Carelessness
On to the next page, starting paragraph five
BEFORE WE ARE MAKING ANY PREMATURE CONCLUSIONS
...
JUST STOPPING
Something is missing
GO AND LEAVING THEM.
Problems with the –ing-form
Unclear paragraph
MANY CHILD AND THEIR FAMILY WILL STARVE AND POSSIBLE

perhaps POSSIBLY
...
Paragraph six, the last paragraph
INSTEAD OF STOP, should be STOPPING, underlining.
And the last line: subject-verb agreement error:
THAT HAVE TO BE TURNED.

Phase 3:I am looking back and taking an overview of the whole essay
Considering 6, perhaps 7, but 7 is probably too much, if I look at the serious grammatical errors, stupid spelling errors.
My assessment of the essay would be a high 6.
It is acceptable, but not average
Language: too many errors to be a 7, a high 6.
Finished with essay 1.
(Ken, script 1)

4.3.1 Phase 1

15 out of 20 raters[45] went through Phase 1, in which the raters would form their initial impressions by scanning the length of the script and/or assess the title, as can be seen above from Ken's rating of script 1. Sometimes this initial phase included an assessment of the overall organizational structure as well, as illustrated by Jens, script 4:

Så er det opgave 4. når jeg kigger ned over den, er det en opgave der knap holder længden. Den er delt ind i et passende antal afsnit, så den ser umiddelbart struktureret ud. Der er også en overskrift CHEAP LABOUR – THE FLIPSIDE OF GLOBALISATION? Og det virker, det virker sådan lidt kreativt.
(Jens, script 4)

Translation:
Now to essay 4. When I scan it, it is an essay that is not as long as it is supposed to be. It is divided up into an appropriate amount of paragraphs, so it looks pretty structured. There is also a title CHEAP LABOUR – THE FLIPSIDE OF GLOBALIZATION? And it seems, it seems a bit creative.
(Jens, script 4)

4.3.2 Phase 2

[45] Some of these raters went through this initial phase with every single script whereas others only went through it with some of the scripts.

All raters went through the second phase. Here they would construct a thorough image of the scripts while making some scattered judgments of them. Here the raters mostly employed interpretation strategies, assisting their reading by mostly classifying or editing errors and by discerning or summing up the ideas in the script. They would also occasionally comment on the script's structure and style/format.

These interpretation strategies were often interspersed with judgment strategies, in which the raters would evaluate the different features of the script (as also illustrated in Ken's protocol above). The majority of the raters saved their score articulation to the end of their assessment, although a few of them assigned preliminary scores before having read the entire script:

> I øjeblikket tænker jeg sådan at det er sådan en opgave solidt i midten, en 8-stykker. Det er også det at tage hensyn til at det selvfølgelig er lidt til den korte side, så, så der skal noget til hvis den skal over 8.
> THE ANSWER IS NO. IF IT WAS SO EASY EVERYBODY WOULD BE HAPPY
> (Jens, script 12)
>
> *Translation:*
> *Right now I am thinking that it is an essay placed solidly in the middle, an 8. That it is, of course, short also needs to be taken into consideration, so so it needs more to be more than an 8.*
> *THE ANSWER IS NO. IF IT WAS SO EASY EVERYBODY WOULD BE HAPPY*
> *(Jens, script 12)*

4.3.3 Phase 3

During the final phase, where the assessments were finalized and a preliminary score assigned, the raters would sum up their assessment, revisiting their judgments, often with an emphasis on the amount of grammatical mistakes:

> Det der sådan set de fejl jeg har fundet.
> Ehrm, jeg har, jeg synes altså overskriften var god. Der er lidt over halvanden side. Jeg synes at sproget det ligger sådan mellem almindeligt og rigtig godt. Altså det er pænt sprog, og fejl, der er flere af dem der er gentagelse. Jeg synes det ligger sådan på på middel. Og der er ikke nogen fejl der forringer noget. Jeg synes egentlig det er er logisk rækkefølge i den måde man har skrevet det på i både afsnit og i sætninger. Og der er taget oplysninger fra fra teksterne, så jeg vil give et 8-tal.
> (Pernille, script 5)

Translation:
These are the errors I have found.
Ehrm, I have, I do think that the title was good. It is a little more than a page and a half. I think that the language is somewhere between average and really good. So it is a nice language, and errors, several of them recur. I think that it is about about average. And none of the errors make it worse. I actually think there is a logical order to how it is set up both with respect to paragraphs and with respect to sentences. And there is information from the source texts, so I will give it an 8.
(Pernille, script 5)

Some raters, however, were aware of their tendency to focus on language aspects and so made an effort to not forget other aspects of the script:

Gælder om at tage nogle notater til den fælles evaluering.
'nummer 1', skriver jeg her. Og karakteren vil jeg nok sige ligger mellem, mellem, ja
6, 5.
Der er alvorlige grammatiske fejl og stavefejl.
Og
Hvis vi kigger på indholdet, som man også skal se på, så er der ikke. Nu vil jeg lige prøve at kigge på det her engang til.
...
Jeg går tilbage til side et for at prøve at se på hvad der er af indholdsaspekter der kan trække opgaven op. Problemet er jo at når der er temmelig mange grammatiske fejl, så hæmmer det læsehastigheden og hæmmer at få fat i indholdet.
(Tina, script 1)

Translation:
It is all about making notes in preparation for the communal rating session I am writing 'number 1' here. And the score, I would say between, between, yes 6 and 5.
There are severe grammatical errors and spelling errors.
And
If we take a look at the content, which also needs to be considered, then there aren't. Now I am just going to try to look at it one more time.
...
I am going back to page one to try and see what content-related features could raise the score. The problem is that when there are quite a lot of grammatical errors, then it obstructs the speed of reading and it is difficult to concentrate on the content.
(Tina, script 1)

4.3.4 Summary of the Sequence of Raters' Decision-Making Behaviors in Independent Ratings

A synopsis of the prototypical macro-level sequence of decision-making behaviors appears in Figure 4.1.

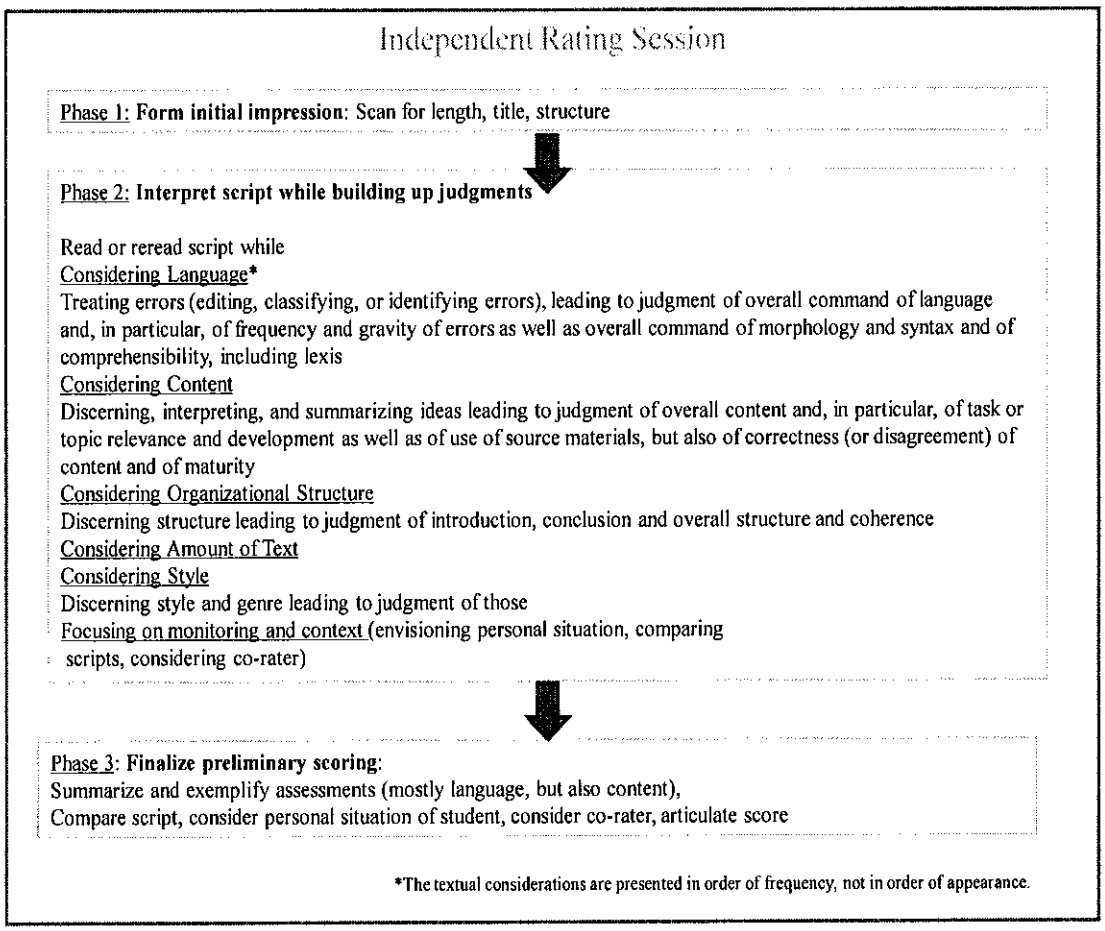


Figure 4.1 Prototypical Sequence of Raters' Decision-Making Behaviors in Independent Rating Sessions

4.4 Balance of Attention to Official Assessment Criteria

The division of the coding system into major categories corresponding to the official assessment criteria allowed me to obtain an account of how the raters distributed their attention across the official criteria, or at least to the textual features corresponding to the official criteria. Although

the raters had different ways of verbalizing their assessments, it was clear from the data that their comments related to the official criteria. The official criteria for the exam in this study were:

- Amount of Text
- Organizational Structure
- Content and Use of Source Materials
- Language
- Style and Format

As emphasized by Lumley (2005), what can be discerned is only the raters' *observable* distribution of attention to the textual features in the scripts because "think-aloud (TA) protocols allow us access only to the descriptions of thoughts and behaviors which the raters articulate" (2005:45).

Figure 4.2 illustrates the raters' distribution of attention to the textual features corresponding to the official assessment criteria. See Table 4.2 (above) for the Mean Percentages and Standard Deviations of Textual Focus in the independent ratings.

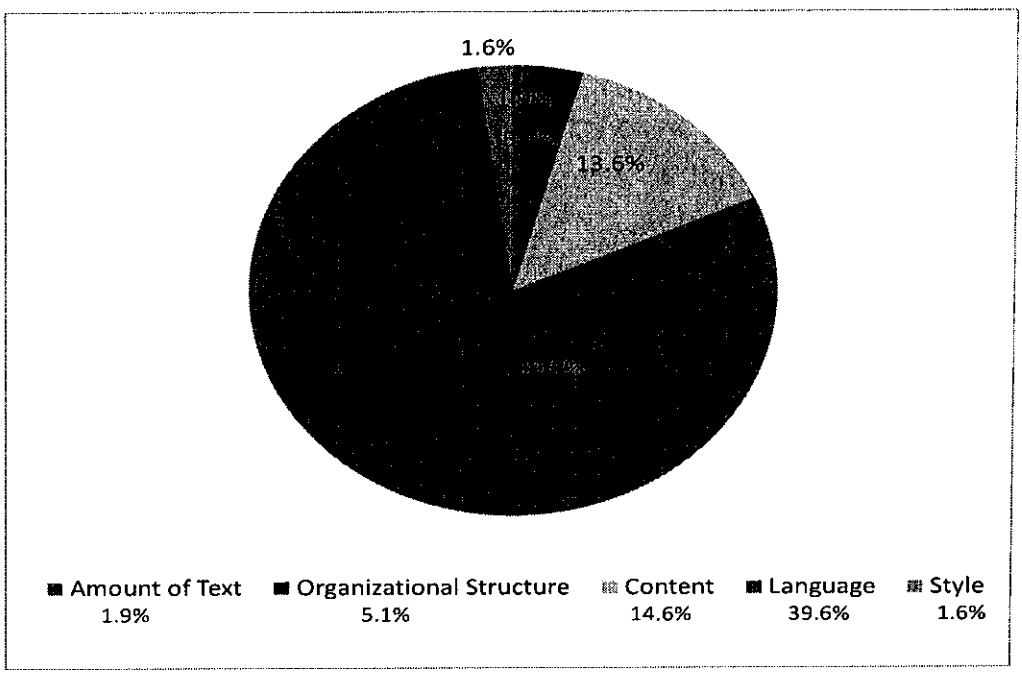


Figure 4.2: Mean Percentages of Textual Features Attended to in Independent Ratings

4.4.1 Interpreting and Judging: a Progression?

The distinction made in the coding scheme between interpretation strategies and judgment strategies facilitated an examination of the raters' distribution of attention to the textual features relative to when they attempted to create an image of the scripts (interpretation strategies) and to when they judged that image (judgment strategies).

Figure 4.3 shows the raters' distribution of attention to the textual features corresponding to the official criteria when, respectively, interpreting and judging the student scripts.

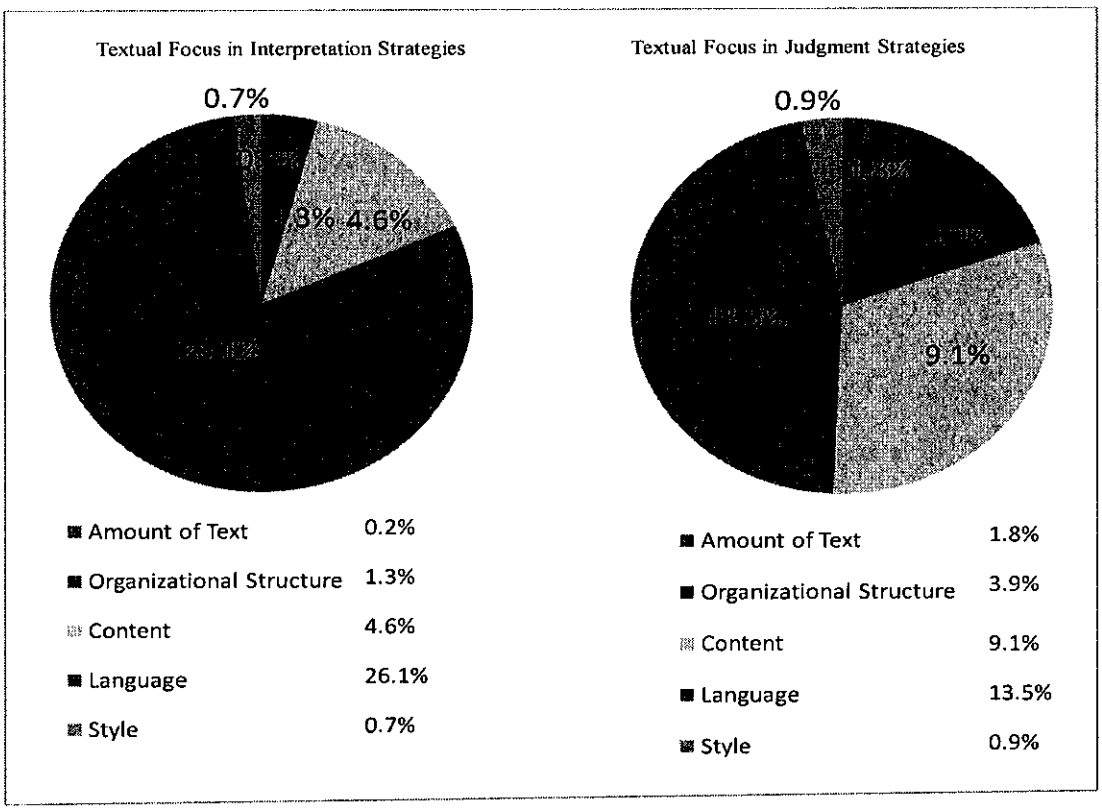


Figure 4.3: Mean Percentages of Textual Features Attended to when Interpreting and Judging Student Scripts

As Figure 4.3 illustrates, the raters' distribution of Textual Focus differed when they interpreted the student scripts and when they judged them. When interpreting the student scripts, the raters

concentrated on language-related features: In 26.1% of their total amount of decision-making behaviors the raters focused on language-related features when interpreting the scripts. This language focus was reduced to almost half when the raters judged the scripts (13.5%), leaving more room for attention to other textual features such as Content and Use of Source Materials (which increased from 4.6% while interpreting to 9.1% while judging), Organizational Structure (increased from 1.2% to 3.9%) and Style and Format (increased from 0.2% to 1.8%).

4.4.2 Summary of Raters' Balance of Attention to Official Assessment Criteria

All of the assessment criteria in the official rubric (Amount of Text, Organizational Structure, Content and Use of Source Materials, Language, and Style/Format) seemed to be taken into consideration when the raters assessed the student scripts in the independent rating sessions. The textual features related to language seemed to attract the most attention, leaving less room for other textual features of the scripts. However, when textual focus was split into interpretation strategies and judgment strategies, it became apparent that the raters focused predominantly on language when creating their image of the student scripts, whereas this intense focus reduced distinctly when the raters judged their script images, freeing up space for attention to the other textual features.

4.5 Rater Agreement

As can be seen from Table 4.4 below, Kendall's W showed an agreement of .87 among the raters for the 300 student scripts scored (1 indicates full agreement and 0 indicates no agreement at all).

Table 4.3: Kendall's W in Independent Ratings

N	20
Kendall's *W*	.87
Chi-Square	243.25
Df	14
Asymp. Sig.	.000

Cronbach's Alpha showed an agreement level of $\alpha = .75$. This means that the raters in this study were in high agreement when it comes to rank ordering the scripts, but they exhibited some degree of variance in the scores they assigned to the scripts.

4.6 Summary of Raters' Decision-Making Behaviors in Independent Ratings

The raters in this study displayed a complex and multi-faceted pattern of decision-making behaviors in independent rating sessions. The raters typically engaged in a rating sequence consisting of three phases (1: form initial impression, 2: interpret while building up judgments, and 3: finalize preliminary scoring). In creating an image of the student scripts (through interpretation strategies) and judging these constructed images (with the use of judgment strategies) the raters would attend to different features of the scripts, corresponding to the official assessment criteria. Although the raters seemed to concentrate their attention on language-related features of the scripts, especially language errors, when interpreting the scripts, they reduced this focus considerably when judging the scripts, thereby welcoming a focus on other textual features. In addition to the textual focus the raters would rely on contextual features (such as other scripts, exam situation, and co-raters) and would monitor and deliberate on their own rating behaviors in interpreting and judging the scripts. No rater ever once consulted the scale descriptors directly. Instead, the raters put their faith in the succeeding communal rating sessions to validate their strategies and finalize their scores. They even prepared themselves for the negotiations in the communal rating sessions by noting specific and concrete examples of the student scripts. The agreement level of the raters' independent scores were .75 (Kendall's W) and .87 (Cronbach's Alpha).

Chapter 5

Raters' Decision-Making Behaviors in Communal Rating Sessions

5.1 Purpose and Scope of the Chapter

This chapter presents results on the raters' decision-making behaviors during the communal rating sessions, using the same coding scheme and approach as in Chapter 4, which reported the results on the independent ratings. The scripts and the raters are the same as in the independent ratings. The data, however, were not think-aloud protocols but recordings of spoken discourse between raters in the rater dyads of the communal rating sessions. Thus, this chapter reports the decision-making behaviors of the same raters rating the same student scripts as in the independent rating sessions, but whereas chapter 4 reported on the raters' decision-making behaviors when they rated student scripts on their own, this chapter reports on the raters' decision-making behaviors when they rated the same scripts but in conversations with another rater.

The chapter is organized around the purpose of addressing the research questions pertaining to communal rating sessions. Most of the sections correspond to the sections in Chapter 4, which reported on raters' independent rating behaviors, although a section is added here that reports on the level of equal engagement in the rater dyads. First, raters' distinct decision-making behaviors are presented by frequency counts interspersed with exemplars of their behaviors. This presentation of their trends is organized according to the raters' focus in their interpretation strategies and their focus in their judgment strategies. Next, prototypical sequences of the raters' behaviors are presented as well as the extent to which their textual foci corresponded to the official assessment criteria. Particular to this chapter of the findings, a report is presented on the extent to which the raters engaged equally in the communal rating discussions, with a focus on

the level of score dominance and the level of conversational dominance. Finally, the level of agreement among the rater dyads are reported.

5.2 Raters' Distinct Decision-Making Behaviors

As was the case when presenting raters' decision-making behaviors in the independent rating sessions, the raters' decision-making behaviors in the communal ratings here are presented as frequency counts interspersed with illustrative examples of the raters' typical behaviors.

Table 5.1 shows the frequency of raters' decision-making behaviors identified during the communal rating sessions of the present study. Coded data were converted to percentages of the total number of decisions made by the individual raters. Table 5.2 provides the same findings as Table 5.1, but in an aggregated form corresponding to the main categories of the coding scheme.

Table 5.1: Mean Percentages (M) and Standard Deviations (SD) of Raters' Decision Making Behaviors in Communal Rating Sessions

***N* = 20 Raters rating 15 student scripts = 300 protocols**	***M(SD)***
Interpretation Strategies	
Contextual and Monitoring Focus	
Read or interpret task input/source materials	0.2% (0.8)
Read or reread student script	0.0% (0.0)
Envision personal situation of the student	2.9% (6.2)
Consider task or exam requirements	1.2% (3.1)
Consider own perception of correct English (e.g. consult a dictionary)	0.1% (0.4)
Textual Focus (Amount of Text)	
Scan script for length	0.0% (0.0)
Textual Focus (Organizational Structure)	
Discern or scan for organizational structure	0.0% (0.0)
Textual Focus (Content and Use of Source Materials)	
Discern or summarize ideas	0.1% (0.5)
Identify or interpret ambiguous or unclear phrases	0.6% (1.6)
Textual Focus (Language)	
Classify language errors into types	0.0% (0.0)
Identify errors	0.1% (0.2)
Correct or edit language (errors or unclear phrases)	0.4% (1.1)

Textual Focus (Style and Format)	
Discern style, register or genre	0.1% (0.2%)
Judgment Strategies	
Contextual and Monitoring Focus	
Articulate score	21.7% (12.5)
Compare student script	2.8% (5.2)
Define, revise or suggest assessment strategies	6,6% (8.3)
Articulate general impression	3.3% (5.4)
Deliberate/articulate teaching strategies/practices	0.7% (2.2)
Exemplify directly from student script	7.4% (8.7)
Consider consensus-based strategy	0.5% (1.4)
Consider personal response or bias	0.3% (1.0)
Textual Focus (Amount of Text)	
Assess or justify amount of text	4.1% (6.5)
Textual Focus (Organizational Structure)	
Assess or justify organizational structure overall	1.4% (3.2)
Assess or justify title	2.3% (5.0)
Assess or justify introduction and/or conclusion	1.8% (4.0)
Assess or justify coherence and/or cohesion	2.1% (4.0)
Textual Focus (Content and Use of Source Materials)	
Assess or justify content/ideas overall	3.5% (5.8)
Assess or justify reasoning, logic, or topic development	3.6% (6.2)
Assess or justify clarity	2.4% (4.4)
Assess or justify correctness of or disagreement with content	1.0% (2.3)
Assess or justify maturity or independence	2.6% (4.9)
Assess or justify task/topic relevance or completion	3.5% (5.8)
Assess or justify use and understanding of source material	3.0% (5.3)
Textual Focus (Language)	
Assess or justify language overall	4.0% (6.6)
Assess or justify frequency of errors	3.6% (6.2)
Assess or justify gravity of errors	1.1% (2.6)
Assess or justify syntax or morphology	2.1% (4.6)
Assess or justify lexis	0.6% (2.1)
Assess or justify fluency or comprehensibility	3.1% (6.1)
Assess or justify spelling	1.1% (2.8)
Assess or justify punctuation	0.0% (0.0)
Textual Focus (Style and Format	
Assess or justify style	2.5% (5.7)
Assess or justify genre	1.6% (3.8)

Table 5.2: Grand Mean Percentages (M) and Standard Deviation (SD) of Raters' Decision Making Behaviors in Communal Rating Sessions

20 Raters rating 15 scripts each = 300 protocols	Contextual and Monitoring Focus	Textual Focus					*Total*
		Amount of Text	Organizational Structure	Content and Use of Source Material	Language	Style and Format	
Interpretation Strategies	4.5% (1.9)	0.0% (0.0)	0.0% (0.0)	0.7% (1.3)	0.5% (1.1)	0.1% (0.2)	*5.7% (3.5)*
Judgment Strategies	43.2% (8.7)	4.1% (2.7)	7.7% (3.0)	19.6% (6.2)	15.7% (3.9)	4.1% (2.3)	*94.3% (3.5)*
Total	*47.6% (8.7)*	*4.1% (2.7)*	*7.7% (3.0)*	*20.3% (6.4)*	*16.2% (3.7)*	*4.1% (2.3)*	*100%*

5.2.1 Interpreting the Student Scripts

Not surprisingly, as the raters had already interpreted the student scripts during the independent rating sessions, the communal rating sessions were dominated by judgment strategies. The communal rating sessions were, however, not completely devoid of interpretation strategies. An average of 5.7% of the strategies expressed during the communal rating session were interpretation strategies.

At times the raters would monitor themselves (or their co-rater) by Envisioning the Personal Situation of the Students (M=2.9%), for example, by imagining the gender of the students:

> *Hans*: der er lidt følelser og det mener jeg også kommer til sidst. Det kan så være det er en pige der har skrevet det.
> *Henrik*: ja, det har jeg på fornemmelsen. Ja, det tror jeg også.
> (Hans and Henrik, script 6)
>
> *Translation:*
> *Hans: there are some sentimental feelings, and I think we also see them towards the end. It could be that it is a girl.*
> *Henrik: yes, I had the feeling. I think so too.*
> *(Hans and Henrik, script 6)*

Other times, the raters envisioned how students handled the time constraints while they were sitting for the exam:

> *Jens*: men den virker sådan ufærdig. Den virker som om vedkommende, ja det var hvad jeg kunne nå. Altså tiden er gået.
> *Nina*: ja, det er rigtigt, det er rigtigt
> *Jens*: altså uden at jeg selvfølgelig kan bevise det. Og så er den jo altså til den korte side, netop af den grund, ikk'?
> *Nina*: ja, det er den.
> (Jens and Nina, script 13)
>
> *Translation:*
> *Jens: but it somehow seems unfinished. It seems as if the test taker, yes, what did I have time for. Time is up.*
> *Nina: yes, it is true, it is true*
> *Jens: well without being able to prove it. And it is on the shorter side, exactly for that reason, right"*
> *Nina: yes, it is.*
> *(Jens and Nina, script 13)*

Similarly, they made interpretations about the students' background knowledge:

> *Torben*: men jeg har da også indtrykket af at de her elever har haft noget omkring det. De inddrager bl.a. også noget sagen omkring McDonald's.
> *Tina*: ja
> *Torben*: og sådan noget. Det er nok ikke noget de sådan bare har. Det er nok fordi de har læst nogle tekster.
> *Tina*: ja, altså. Det virker da. Ja, heldigvis da, som om de har arbejdet med det emne i forvejen.
> (Toben and Tina, script 1)
>
> *Translation:*
> *Torben: but then I also get the feeling that that these students have worked with this before. Among other things, they also include the case about McDonald's.*
> *Tina: yes*
> *Torben: and such things. It is not just something they know. It is probably because they have read some texts.*
> *Tina: yes, really. It seems as if. Yes, fortunately, as if they have worked with this topic before.*
> *(Torben and Tina, script 1)*

The raters would also at times Consider Task or Exam Requirements (M=1.2%). This even led, in some cases, to clarification of the task requirements:

> *Nina*: ja, også, jeg kan ikke huske om den her har noget med den danske. Det har jeg også kigget på i hvert fald om de har fat, for det er et krav om at de skal anlægge

Jens: hmm, nej det er det ikke.
Nina: jo det er.
Jens: nå, er det det. Står der ikke bare "kan inddrages"?
Nina: I BESVARELSEN INDDRAGES … DEN DANSKE OG DEN ENGELSKE
Jens: ok, yes.
(Nina and Jens, script 2)

Translation:
Nina: yes, that also, I don't remember whether they have included something from the Danish text. I also take into consideration whether they have, because they are required to.
Jens: hmm. No it is not.
Nina: yes it is.
Jens: well, isn't it. Doesn't it just say "can include?"
Nina: IN THE ESSAY ... THE DANISH AND THE ENGLISH TEXT MUST
Jens: ok, yes
(Nina and Jens, script 2)

Clarification was, in fact, the purpose of most of the talk among the dyads of raters related to interpreting the scripts. Often this took the form of interpreting ambiguous or unclear phrases (Identify or Interpret Ambiguous or Unclear Phrases: M=0.6%):

Pernille: men jeg har også rigtig mange fejl. Og jeg har bl.a. også ude for et afsnit skrevet 'hvad snakker du om?', fordi det kunne jeg slet ikke finde ud af.
Jesper: hvad var det for et?
Pernille: Jammen, det er på anden side, det øverste.
Jesper: ja
Pernille: ehrm
Jesper: jo, det tror jeg nok efterhånden jeg fandt ud af.
Pernille: Ha, ha, der var du meget involveret.
Jesper: Ja, nej, nu skal du bare høre. Det er noget med Adidas og Nike, de kom øverst på, på den der undersøgelse, fordi de frivilligt besvarede på spørgsmålene. Reebok og New Balance besvarede kun nogle få af spørgsmålene. Fila har overhovedet ikke besvaret. Og den der undersøgelse, den omfattede kun virksomheder der producerer løbesko. Men det kunne ligeså godt havce været et hvilket som helst andet virksomhedsprodukt.
Pernille: ja
Jesper: Er det ikke. Tror du ikke det er sådan?
Pernille: Jammen, hvad siger du til den der THEY DIDN'T ANSWER FEW OF THE QUESTIONS.
Jesper: det betyder da bare at der var nogle få spørgsmål de ikke besvarede.
Pernille: Ja, det er selvfølgelig rigtig nok.
(Pernille and Jesper, script 15)

Translation:
Pernille: but I also have a lot of, a lot of errors. And I also wrote next to a paragraph "what are you talking about?" because I simply didn't know.
Jesper: what paragraph was it?

Pernille: well, it is on page two, the top one.
Jesper: yes
Pernille: ehrm
Jesper: yes, I think that I found out along the way.
Pernille: Ha, ha, you were very involved
Jesper: Yes, no, now let me tell you. It is something about Adidas and Nike, they came out on top in this survey because they volunteered to answer the questions. Reebok and New Balance answered but few of the questions. Fila didn't answer at all. And this survey, it only focused on companies that produce running shoes. But it might as well have been any other product.
Pernille: yes
Jesper: isn't it. Don't you think it is like that?
Pernille: yes, but what do you say to this one THEY DIDN'T ANSWER FEW OF THE QUESTIONS.
Jesper: it just means that there were few questions that they did not answer.
Pernille: yes, of course. You are right.
(Pernille and Jesper, script 15)

Such attempts to interpret the content of the students' scripts also related to their language, for example, resulting in editing unclear phrases: (Correct or Edit Language Errors or Unclear Phrases): *M*=0.4%):

Tina: jo, altså
Torben: TO SIGN ON A CONTRACT;
Tina: ja, det der med ON. Det er bare TO SIGN A CONTRACT.
Torben: men der er jo ikke deres ansatte. Det er vel netop deres selskaber, at de skal skrive en kontrakt på at de vil overholde de her regler.
Tina: Ja, men det kan måske vare lidt svært at se, hvad Reeboks rolle er i forhold til deres, om det er datterselskaber eller om det er leverandører.
...
Torben: jo, jo. Men det er jo det den går på, ikke? At de vestlige firmaer har svært ved at kontrollere de der suppliers de har ude i tredieverdens lande.
Tina: jo
Torben: hvor de har lagt produktionen
Tina: men altså, spørgsmålet er så om det er neutrale leverandører, om jeg så må sige, eller om de er ansat af Reebok. Altså, de
Torben: ja, ja
Tina: jo, men synes du så ikke stadigvæk at EMPLOYERS er et forkert ord så?
(Torben and Tina, script 8)

Translation:
Tina: yes, so
Torben: TO SIGN ON A CONTRACT;
Tina: yes, this one ON. It is just TO SIGN A CONTRACT.
Torben: but it isn't their employees. On the contrary, it is their companies that they have to sign a contract that they are going to obey the rules.

Tina: Yes, but it might be a bit difficult to see what role Reebok plays in relation to their, whether they are subsidiary companies or whether they are contractors.
...
Torben: yes, yes. But that's the line it takes, isn't it? That the western companies have difficulties in controlling their contractors in the third world.
Tina: yes
Torben: where they have placed their production
Tina: but really, the question is whether the contractors are neutral or whether, if I may, or whether they are employed by Reebok. Well, they
Torben: yes, yes
Tina: Yes, but don't you still think that EMPLOYERS is the wrong word then?
(Torben and Tina, script 8)

In some instances, aspects of the student scripts came to light that had escaped the attention of one of the raters during the person's independent rating (Classify errors and Identify errors: *M*=0.5%).

Malene: og jeg kan se der adjektiv og adverbier og der er
Thea: ja, det har jeg noteret i hvert fald
Malene: kongruensfejl, ikke? Og der er stavefejl.
Thea: præpositioner.
Malene: ja, og præpositioner er der også, ja. Dem har jeg så ikke noget tal for, men ehrm det er rigtigt at det er galt med. Så det er det er, der har jeg, der har jeg. Der tror jeg, jeg har sovet.
Thea: ja
Malene: simpelthen. Men ehrm spørgsmålet er, der må varere noget med, der må være noget med jeg har syntes at at indholdet i hvert fald har været fornuftigt.
Thea: ja,
(Malene and Thea, script 5)

Translation:
Malene: and I can see that there are adjectives and adverbs and there are
Thea: yes, that at least I have noticed.
Malene: subject-verb agreement, right? And there are spelling errors.
Thea: prepositions.
Malene: yes, and there are prepositions, too, yes. I don't have a number for those, but ehrm, it is true that something is wrong about them, too. So that is why, there I have, there I have. It must have escaped my attention.
Thea: yes
Malene: just like that. But ehrm, the question is, it must be something, it must be because I thought the content, at least, made sense.
Thea: yes
(Malene and Thea, script 5)

In a case of high discrepancy (i.e., where the two raters were more than one score apart) one rater even felt the need to reread the script:

Jesper: godt, så er det nummer 2. og det er den der hedder WORK ETHICS til overskrift.
Pernille: der er halvanden side, ja.
Jesper: ja
Pernille: altså, den har jeg givet et 8-tal for.
Jesper: så er vi på den, Pernille.
Pernille: nå
Jesper: jeg har 5 til 6.
Pernille: er det rigtigt?
Jesper: ja, ha, ha
Pernille: ha, ha. åh Gud. Så er jeg næsten nødt til at læse den igennem. Skal jeg ikke?
(Pernille and Jesper, script 2)

Translation:
Jesper: well, number 2. And it is the one with WORK ETHICS as the title.
Pernille: it is a page and a half, yes.
Jesper: yes
Pernille: actually I have given it an 8.
Jesper: so we've got a situation, Pernille.
Pernille: oh.
Jesper: I've got 5 to 6.
Pernille: is that right?
Jesper: yes, ha, ha.
Pernille: ha, ha, oh my God. Then I'll have to read it again. Shouldn't I?
(Pernille and Jesper, script 2)

5.2.1.1 Summary of Raters' Interpretation Strategies in Communal Ratings

During the communal ratings, the raters generally spent little energy creating an image of the student scripts because they had already done so in their independent rating sessions. However, at times they felt it necessary to test their hypotheses about the images they had created earlier during the independent rating sessions. To do this, they would interpret and edit unclear phrases from the student scripts, and they would reconsider contextual factors such as their vision of the student's personal situation as well as task and exam requirements. One rater (Pernille) even felt the need to reread the entire script in cases of strong score disagreement.

5.2.2 Judging the Student Scripts

Although the raters at times felt the need to revisit their images of the student scripts, they spent most of their energy judging these images.

5.2.2.1 Assigning Scores

The raters were all aware that they had to reach a final score by the end of the communal rating session, and so a lot of their talk centered on determining the score (Articulate Score: M=21.7%). In cases of agreement between the raters, the final score was typically the only focus of their discussions:

Jesper: så er det 5'eren. BUSINESS ETHICS IN THE THIRD WORLD
Pernille: ja
Jesper: jeg har skrevet 8 for den.
Pernille: 8?
Jesper: rent
Pernille: ja, det har jeg også
Jesper: fuldstændig enig. Der er
Pernille: ja
Jesper: jeg var lige ved at sige at det var sådan en man man næsten kunne tage og ud og sige: der er middelbesvarelsen efter mine begreber.
Pernille: ja
Jesper: ehrm
Pernille: skal vi gå videre?
Jesper: ja, det synes jeg, fordi der er ikke nogen grund til at
(Pernille and Jesper, essay 5)

Translation:
Jesper: Then we have number 5. BUSINESS ETHICS IN THE THIRD WORLD
Pernille: yes
Jesper: I have written an 8.
Pernille: 8?
Jesper: a clear 8
Pernille: yes, me too.
Jesper: completely agree. There are
Pernille: yes
Jesper: I was just about to say that with an essay like this one, you could just about take it out and say, this is an average essay according to me
Pernille: yes
Jesper: ehrm
Pernille: shall we go on?
Jesper: yes, because there is no use in.

(Pernille and Jesper, script 5)

In cases of score discrepancy, the raters always validated and justified their scores. When the raters were more than one score apart, the raters would take great pains to determine what they believed was the most appropriate score:

Jesper: godt, så er det nummer 2. og det er den der hedder WORK ETHICS til overskrift.
Pernille: der er halvanden side, ja.
Jesper: ja
Pernille: altså, den har jeg givet et 8-tal for.
Jesper: så er vi på den, Grethe.
Pernille: nå
Jesper: jeg har 5 til 6.
Pernille: er det rigtigt?
Jesper: ja, ha, ha
Pernille: ha, ha. åh Gud. Så er jeg næsten nødt til at læse den igennem. Skal jeg ikke?
Jesper: ja, det ved jeg ikke altså. Lad os lige prøve at se hvad, hvad, kigge på vores noter hver især, og så måske lige kigge ned over vores rettelser
Pernille: jeg har skrevet, der er mange fejl. Det er rigtig nok. Og så er den for kort, jo.
Jesper: ja, det er de jo næsten alle sammen.
Pernille: ja, jammen det har jeg også indtalt med at det er de allesammen.
Jesper: altså, jeg har skrevet at på sprogsiden, der synes jeg der er manglende forståelse af, hvad hedder det hovedsætning – bisætningproblematikken. Der er meget lidt forståelse for kongruens. Ehrm, der er problemer med idiomatik. Ehrm, så er den rimelig struktureret.
Pernille: ja
Jesper: ja, og så ja det ved jeg ikke. Så er jeg nok eksploderet lidt over den der slutning BUT LET'S HOPE THAT THINGS WILL CHANGE FOR THE BETTER IN THE YEARS TO COME. Det kan man altså sige om hvad som helst.
Pernille: ja, det er selvfølgelig rigtig nok. nej, det er jeg bedyret over.
Jesper: hvad, hvad har du skrevet op af godt og skidt?
Pernille: jammen, jeg har også skrevet det der. jeg synes egentlig det var en meget god ehrm frame. Altså det der bygget op, den måde det er bygget op på. Du brugte et andet ord. Ehrm, men jeg har da også nogle kongruensfejl men så mange har jeg ikke. Jeg har 5. Er der mange flere?
Jesper: 2,3,4,5,6,7, 8. jeg har 9.
Pernille: nå, jammen så er det mig der ikke kan min Grammatik.
Jesper: ha
Pernille: ha, jeg har også kun læst den igennem een gang.
Jesper: det er jo heller sådan nødvendigvis at tallet der
Pernille: nej, nej, men alligevel. Ahhh.
...
Jesper: næh, altså, jeg, jeg. Det har jeg ikke.
...
Pernille: altså, jeg er nødt til at læse den igennem igen. Det vil jeg altså meget gerne. Kan vi ikke slukke?
Jesper: jeg tror at vi aftaler at vi holder pause mens vi lige læser den, fordi der vil så ikke være noget snak. Og så genoptager vi snakken.

Pernille: ja
Jesper: så vi pauser.
...
Jesper: ja, vi er på og genoptager diskussionen. Pernille, du sagde
Pernille: ja, jeg vil sige at jeg synes den er bedre end til et 6-tal, fordi jeg synes den er pænt bygget op. og jeg ved godt at der er nogle fejl. Og jeg synes også at kongruensfejl er slemme. Og der er også mange ting. Men jeg synes ikke det er sådan at det ødeligger opstillingen. Jeg synes, jeg kan godt lide den rækkefølge. Jeg kan gode lide de ting, han får med i det, eller vedkommende får med i det, ikke også?
Jesper: ja
Pernille: jeg synes også nogle gange vedkommende får nogle ting sagt meget hurtigt, som andre bruger mere tid på, f.x. der på side to, at Reebok får også andre firmaer med, eller signalerer i hvert fald det her til de andre firmaer plus de har de her awards
Jesper: mm
Pernille: og der jeg synes, jeg synes egentlig. Ja, det kan godt ske at 8 er for meget. Jammen, det er det sikkert. Jeg er sådan bange for at være for skrap. Men men jeg synes også at 6 er for lidt til denne her.
Jesper: jammen, altså
Pernille: ja, det ved jeg ikke. Hvad siger du efter at du har læst den igen?
Jesper: ja, jeg kunne ikke rigtig afgøre med mig selv om, ha. om jeg havde været for skrap. Der er nok også tendens til at når man først har lagt sig fast på en eller anden karakter
Pernille: ja
Jesper: så leder man efter tegn på at det skal være rigtigt, ikke?
Pernille: ved du hvad?
Jesper: nu faldt jeg over tredie afsnit, det, der begynder man at lave referat af, af teksten, ikke? Og så leder man efter de der. hvis du ser omme på anden side, slutning HAS ACHIEVED TO REACH SUCCES IN THE FINANCIAL ASPECT AS WELL AS IN HUMANITY IN GENERAL. Ehrm, der har jeg altså skrevet et stor spørgsmålstegn. Jeg ved godt hvad der menes, men ehrm
...
Pernille: ja
Jesper: men altså
Pernille: der tænker man selvfølgelig jo på det der med
Jesper: ja
Pernille: med at de netop går ind og reagerer
Jesper: ja, ja
Pernille: jeg ser nu altså også en fejl før, ha. ha. så jeg har faktisk også flere. Jeg har 8.
Jesper: ja fordi. Pernille, skal
Pernille: jammen, nej
Jesper: kan du gå med på 7?
Pernille: ja
Jesper: så, ehrm. vi skal da heller fået gjort dig til slagterforhandler
Pernille: jammen, du er mere til et 6-tal.
Jesper: jammen, der er. Ja. men altså, jeg vil sige det på den måde at at jeg synes den er under middel, så ehrm om det bliver 6 eller 7, det er såmænd det, det er et religiøst spørgsmål.
Pernille: hrm
Jesper: der er det, så synes jeg, jeg synes bare den er under.
Pernille: og det er også 7, hvis vi siger at middel er 7.8.9?
Jesper: ja
Pernille: ja

Jesper: altså i den lave ende i hvert fald.
Pernille: ja
Jesper: ja, ja, ja
Pernille: jammen, det, det kan jeg også godt se at det er rigtig nok. Men ehrm, nej erhm, jeg synes at ikke den er til 6. det må jeg sige.
Jesper: nej, jammen så synes jeg vi skal sige
Pernille: skal vi sige?
...
Jesper: godt
(Pernille and Jesper, script 2)

Translation:
Jesper: well, number 2. And it is the one with WORK ETHICS as the title.
Pernille: it is a page and a half, yes
Jesper: yes
Pernille: actually, I have given it an 8.
Jesper: so, we've got a situation, Pernille.
Pernille: oh
Jesper: I've got 5 to 6.
Pernille: is that right?
Jesper: yes, ha, ha
Pernille: ha, ha. Oh my God. Then I'll have to read it again. Shouldn't I?
Jesper: Yes, I don't know about that. Let's see what, what, let's each look at our notes, and then perhaps scan our corrections.
Pernille: I have written that there are many errors. That's true. And then it is, you know, too short.
Jesper: Yes, almost all of them are.
Pernille: yes, but I have also mentioned that.
Jesper: well, I have written about the language that there is no knowledge of, what is it called, the problem of superordinate clauses and a subordinate clauses. Little understanding of subject-verb agreement. Ehrm, problems with idiomatic phrases. Ehrm the structure is ok.
Pernille: yes.
Jesper: yes, and so yes I don't know. I probably exploded because of this ending BUT LET'S HOPE THAT THINGS WILL CHANGE FOR THE BETTER IN THE YEARS TO COME. Well, you can say that about anything.
Pernille: yes, of course, that is true. No, I acknowledge that.
Jesper: what have you written about its strengths and weaknesses?
Pernille: well, I have written the same that. I actually think it was a pretty good ehrm frame. I mean, built up, the way that it was built up. You used another word. Ehrm, but I also do have some subject-verb agreement errors, but not that many. I have 5. Are there more?
Jesper: 2,3,4,5,6,7, 8. I have 9.
Pernille: oh well, then I guess that I don't know my grammar.
Jesper: ha
Pernille: ha, also, I only read it once.
Jesper: well, it is not that the number is.
Pernille: no, no, but anyway. Ahhh.
...
Jesper: no, well, I, I. I don't have.
...
Pernille: really, I have to reread it. I really want to. Can't we turn it off?

Jesper: I think that we agree that we take a break while we are reading it, because no one will be saying anything. And then we resume the talk.
Pernille: yes
Jesper: so we are pausing.
...
Jesper: yes, we are on and are resuming the discussion. Pernille, you said.
Pernille: yes, I would say that it is better than a 6, because I think it has a nice structure, and I know that there are some errors. And I also think that subject-verb agreement errors are bad. And there are so many things. But I don't think that it obstructs the structure. I think I like the order. I like the things he includes or the student includes, right?
Jesper: yes.
Pernille: I also sometimes think that the student says some things very efficiently, things that other people spender more time on, for instance a page or two, that Reebok brings along other companies, or at least signals to the other companies that they have been honored with these awards.
Jesper: mm
Pernille: and that's why I think, I actually think. Yes, perhaps 8 is too much. But, maybe it is. I am a bit afraid of being too strict. But I also think that 6 is too low for this one. Jesper: but ehrm.
Pernille: yes, I don't know. What do you say after you have reread it?
Jesper: yes, I couldn't really make up my mind about whether I had been too strict. There is perhaps this tendency of when you have set yourself on a score.
Pernille: yes
Jesper: then you look for confirmation, right?
Pernille: do you know what?
Jesper: now I came across the third paragraph, it, here they start making a summary, right? And then they look for these, if you go to the second page, conclusion HAS ACHIEVED TO REACH SUCCES IN THE FINANCIAL ASPECT AS WELL AS IN HUMANITY IN GENERAL. Ehrm, I have actually written a big question mark here. I don't know what they mean, but ehrm
...
Pernille: yes
Jesper: but ehrm
Pernille: here of course they are thinking of
Jesper: yes
Pernille: that they really do react.
Jesper: yes, yes
Pernille: I also see an error before, ha, ha, so I do, in fact, have more. I have 8.
jeg ser nu altså også en fejl før, ha. ha. så jeg har faktisk også flere. Jeg har 8.
Jesper: yes, because, Pernille, do
Pernille: yes, but no.
Jesper: would you agree to a 7?
Pernille: yes
Jesper: so, ehrm. We don't want to turn you into a butcher.
Pernille: yes, but you think a 6 is more appropriate.
Jesper: yes, but there are. Yes, but really, let me put it this way: I think it is below average, so whether it ends up as a 6 or a 7, that really doesn't matter.
Pernille: hrm
Jesper: that is it, then I think, I just think it is below.
Pernille: and 7 is too, if we say that average is 7,8,9?
Jesper: yes
Pernille: yes

Jesper: well, in the lower end, at least.
Pernille: yes
Jesper: yes, yes, yes.
Pernille: But ehrm, it, I can see that that is an appropriate score. But ehrm, no, ehrm, I do not think it is a 6. I gotta say that.
Jesper: no, but then I think we should say
Pernille: should we say?
...
Jesper: good
(Pernille and Jesper, script 2)

As can be seen from this excerpt, the final determination of the score was not a dialectic tug-of-war, but came as a result of genuine attempts to reach what the raters reasoned to be the most appropriate score. When, after mutually justifying their scores, the raters failed to reach a consensus on the most appropriate score, they did not consult the scale descriptors directly for guidance, but rather compromised. If there was any direct reference at all in the process of resolving the score, it was to the rater community:

Malene: og så kan man sige at vi plejer jo at gøre det at vi ehrm så lader det komme eleven til gode, så vi giver den høje karakter
Thea: ja, ja
Malene: så skal vi ikke så lande på det?
Thea: vi lander på 7
(Malene and Thea, Script 1)

Translation:
Malene: and then we could say that we usually do we ehrm then we give the student the benefit of the doubt.
Thea: yes, yes
Malene: so shouldn't we say that?
Thea: we'll give it a 7.
(Malene and Thea, Script 1)

Although it appears that in the local rating community the raters tend to assign the higher of the two adjacent independent scores, it was often also the case that the lower of the independent scores ended up as the final score. Occasionally, though, the discussion led to a final score that was even lower than the lower of the two independent scores (see Appendix H for all scores assigned). This mainly resulted from one rater revising her general assessment strategies (toward harshness) over the course of the session:

Jesper: så kommer nummer 14. ETHICS ABOUT CHILD WORK
Pernille: ja
Jesper: og ja, den er jo ekstrem kort.
Pernille: ja
Jesper: ehrm. Og jeg har faktisk skrevet at her mener jeg at den helt, helt rigtige karakter må være et 4-tal.
Pernille: ha, ha
Jesper: den er virkelig, for mig, et grænseland mellem 3 og 5.
Pernille: ja, ok, men der støder vi ind i samme problem vi havde lige før, hvor jeg har givet et 6-tal. Det er fordi. Jeg har skrevet det er et meget lille 6-tal. Og det er fordi jeg har skrevet, det er et meget lille 6-tal. Og hvorfor har jeg skrevet det? Fordi, jeg måske nok synes at der er en lidt opbygning i. Men jeg har skrevet at det er et meget lille 6-tal.
Jesper: ja
Pernille: skal vi så give et 4-tal?
Jesper: Nej, så synes jeg, vi skal give et 5-tal.
Pernille: Nå, jammen hvad havde du givet?
Jesper: Jeg har bare. Jeg har skrevet 03-5. Så har jeg skrevet lig 4.
Pernille: ja
Jesper: ehrm THESE PRODUCTS ARE MADE BY HUMAN RIGHTS.
Pernille: ja, altså ja. Der er virkelig meget. Jammen, det er rigtigt. Den har jeg også. Sproget, det er meget med.
Jesper: ja, men altså
Pernille: men altså, det er en af dem igen, hvor man kan sige, jammen altså, er det 5 eller er det 3, fordi det, det, det er ligesom man sådan synes at få lagt det der. Som jeg sagde til dig, jeg har nok sådan lidt en afstandstagen fra de der meget lave karakterer, når man trods alt har næsten siddet sådan og. Men det er jo altid noget man kan diskutere jo.
Jesper: ja
Pernille: det er det jo.
Jesper: altså, man kan jo sige, at den her, den har jo heller ikke så meget styr på begreberne, vel, altså? Den tredie verden, og så bliver det IT'S VERY DIFFICULT TO HELP THESE WORLDS
Pernille: ja
Jesper: ehrm
Pernille: ja, det er den med WORLDS hele tiden.
Jesper: ja
Pernille: ja
Jesper: MCDONALDS ARE UNFORTUNATELY NOT THE ONLY COUNTRY
Pernille: ja
Jesper: ehrm
Pernille: det er rigtigt
Jesper: altså det
Pernille: og skriver også kun om child work, ikke?
Jesper: jo
Pernille: og som jeg egentlig også synes er
Jesper: ja, og der er jo, der er jo overhovedet ingen indledning, vel? Overhovedet.
Pernille: nej, ikke spor. Altså, vi kan godt skrive 3, fordi det er jo nok også vigtigt at vi begynder og og få karaktererne noget ned. Jammen, det synes jeg.
Jesper: ja
Pernille: Det vil jeg meget gerne være med til.
Jesper: det
Pernille: jeg skal bare til at tage mig sammen og give nogle 3-taller.

Jesper: ha, ha
Pernille: Og nogle 0'er. Jammen, der egentlig også min indstilling. Så, sommetider når man bliver beskyldt for at være skrap, så synes jeg: nå, jammen, det er jeg måske også, ikke? Og uha, så må jeg være noget venligere.
(Jesper and Pernille, script 14)

Translation
Jesper: then we have number 14. ETHICS ABOUT CHILD WORK
Pernille: yes
Jesper: and yes, it is extremely short.
Pernille: yes
Jesper: ehrm. And I have actually written down that here the most, the most appropriate score would be a 4[46].
Pernille: ha, ha
Jesper: To me this is really in between a 3 and a 5.
Pernille: Yes, ok, but here we encounter the same problem as we had just before where I gave given it a 6. It is because. I have written that it is a very low 6. And it is because I have written that it is a very little 6. And why have I done that? Because I probably think that there is some structure to it. But I have written that it is a very low 6.
Jesper: Yes
Pernille: Should we then give it a 4?
Jesper: No, then I think we should give it a 5.
Pernille: Oh, but what did you give it?
Jesper: I have just. I have written 3 to 5. So I have written equals 4.
Pernille: Yes
Jesper: ehrm THESE PRODUCTS ARE MADE BY HUMAN RIGHTS.
Pernille: Yes, so yes. There really is a lot. But it is true. I have that too. The language, there is a lot of.
Jesper: Yes, but then
Pernille: but so, it is one of these again, where one could say, but really, is it a 5 or is it a 3, because it, it is kind of like the way we kind of put it. As I told you, I probably avoid these very low grades when the student has spent time. But that is a debatable matter.
Jesper: yes
Pernille: indeed it is
Jesper: so, you could say that this one, this one doesn't know about the concepts, right? The third world, and then it becomes IT'S VERY DIFFICULT TO HELP THESE WORLDS
Pernille: yes
Jesper: ehrm
Pernille: yes, it is the thing about the WORLDS all the time.
Jesper: yes
Pernille: yes
Jesper: MCDONALDS ARE UNFORTUNATELY NOT THE ONLY COUNTRY
Pernille: yes
Jesper: ehrm
Pernille: that is true
Jesper: well it
Pernille: and writes only about child work, right?

[46] The score of 4 is non-existent in the scale used to rate the scripts. It is a score between a 3 and a 5.

Jesper: yes
Pernille: and what I actually also think is
Jesper: yes, and there is, there is, in fact no introduction, right? At all.
Pernille: no, not at all. So, we can write a 3, because it also important that we start assigning lower scores. But I think so.
Jesper: yes
Pernille: I would like to go along with that.
Jesper: it
Pernille: I just have to pull myself together and assign some 3s.
Jesper: ha, ha
Pernille: and some 0s. But I really do think so. So, sometimes when I get accused of being too harsh, then I think, then I think: yes well, but, perhaps I am, right? And oh no, I better be nicer.
(Jesper and Pernille, script 14)

5.2.2.2 Justifying Judgments: Monitoring and Contextual Focus

In justifying their scores, the raters would put forth their overall impressions (Articulate General Impression; M=3.3%). More often, however, they would exemplify directly from the student scripts (Exemplify Directly from Student Script: M=7.4%):

Helle: jammen, der er trods alt meget mere indhold i.
Astrid: meget mere, og så har jeg sat rosende stjerner ehrm over for, over for forskellige ting og sager, for eksempel, der er noget her THE CANDIDATES i tredier afsnit; elevens tredie; selvom der er ikke er mellemrum
Helle: mellem de to førstes, ikke?
Astrid: THE CANDIDATES PREPARE SPEECHES TO CONVINCE THE WORKERS
Helle: ja
Astrid: IN THE BATTLE TO WIN AND IMPROVE. Det synes jeg har en retorisk værdi i sig, ikke?
Helle: ja
Astrid: i sig, ikke?
Helle: det er meget flot
Astrid: jeg kan godt lide også at hun i næste afsnit, linie three siger, snakker om DRASTIC MEASURES
Helle: ja, det er rigtigt
Astrid: for eksempel, og TEMPORARY PERIOD
Helle: ja
Astrid: det skal ikke omskrives med noget upræcist.
Helle: nej
Astrid: hun rammer noget og på et plan som jeg synes er godt.
(Astrid and Helle, script 2)

<u>Translation:</u>
Helle: yes, but you could say that the content is better.

Astrid: much better, and then I have placed a nice star ehrm next to, next to various features, for instance, there is something here THE CANDIDATES in the third paragraph, the student's third; even though there is no space between.
Helle: between the two first ones, right?
Astrid: THE CANDIDATES PREPARE SPEECHES TO CONVINCE THE WORKERS
Helle: yes
Astrid: IN THE BATTLE TO WIN AND IMPROVE. I think this has some value, rhetorically, right?
Helle: yes
Astrid: in itself, right?
Helle: it is very beautiful
Astrid: I also like the fact that in the next paragraph, line three, she talks about DRASTIC MEASURES
Helle: yes, that is true
Astrid: for instance and TEMPORARY PERIOD
Helle: yes
Astrid: it does not need to be rewritten with something imprecise.
Helle: no
Astrid: she is touching on something and on a good level, I think.
(Astrid and Helle, script 2)

As in the independent rating sessions, the raters also judged or justified the scripts by comparing them to one another (Compare Scripts: M=2.8%). They acknowledged, though, that they were not supposed to do that:

Jens: og der har jeg måske gået lidt hårdt til den.
Nina: ja, jeg har så plusset den i forhold til de andre. Ja, man må jo ikke sammenligne.
Jens: nej, men det gør man så jo alligevel et eller andet sted, ikke?
Nina: det er svært nogle gange, synes jeg i hvert fald. .. Det at vedkommende har brugt den danske tekst.
Jens: ja, ok
Nina: For eksempel, den sidste vi har givet 9. lad os se, jeg kan ikke huske om. Nu må jeg se. nej, nu må vi ikke sammenligne.
(Nina and Jens, script 5)

Translation:
Jens: and with this one I have perhaps been a bit harsh.
Nina: yes, I have found some good parts compared to the other scripts. Yes, I now we are not supposed to compare.
Jens: no, but we do that anyway, don't we?
Nina: it is difficult some times. At least I think so. The fact that the test taker has used the Danish article.
Jens: yes, ok
Nina: for instance, we gave that one a 9. Let us see, I don't remember whether. Now I gotta see. No, we are not supposed to compare.
(Nina and Jens, script 5)

In justifying their scores the raters often seemed to develop themselves professionally as raters. Some accepted to become a harsher rater, as Pernille did in her interaction with Jesper in script 14 (excerpt reproduced here):

> *Pernille*: nej, ikke spor. Altså, vi kan godt skrive 3, fordi det er jo nok også vigtigt at vi begynder og og få karaktererne noget ned. Jammen, det synes jeg.
> *Jesper*: ja
> *Pernille*: Det vil jeg meget gerne være med til.
> *Jesper*: det
> *Pernille*: jeg skal bare til at tage mig sammen og give nogle 3-taller.
> *Jesper*: ha, ha
> *Pernille*: Og nogle 0'er. Jammen, der egentlig også min indstilling. Så, sommetider når man bliver beskyldt for at være skrap, så synes jeg: nå, jammen, det er jeg måske også, ikke? Og uha, så må jeg være noget venligere.
> (Jesper and Pernille, script 14)
>
> *Translation:*
> *Pernille: no, not at all. So, we can write a 3, because it also important that we start assigning lower scores. But I think so.*
> *Jesper: yes*
> *Pernille: I would like to go along with that.*
> *Jesper: it*
> *Pernille: I just have to pull myself together and assign some 3s.*
> *Jesper: ha, ha*
> *Pernille: And some 0s. But I really do think so. So, sometimes when I get accused of being too harsh, then I think, then I think: yes well, but, perhaps I am, right? And oh no, I better be nicer.*
> *(Jesper and Pernille, script 14)*

Others also deliberated over their assessment strategies (Define, Revise or Suggest Assessment Strategies: M=6.6%), for example, coming to realize what they, unlike others, tend to focus on:

> *Nina*: ehrm, lad os se. I DID NOT KNOW ABOUT REEBOK'S ELECTIONS. Igen bruger den i hvert fald.
> *Jens*: mmhm
> *Nina*: både den engelske og
> *Jens*: mhm
> *Nina*: og det er rigtig nok med sproget. Det er ikke.
> *Jens*: Jammen, det er klart. Jeg har nok fokuseret mere ehrm på sproget
> *Nina*: ja, ja
> *Jens*: i forhold til indholdet her, fordi fordi der er dem der. Der er en del tilfælde hvor stedord bliver brugt uden at det er klart hvad der refereres til. Det irriterer mig grænseløst.
> (Nina and Jens, script 5)

Translation:
Nina: ehrm, let us see. I DID NOT KNOW ABOUT REEBOK'S ELECTIONS. Again it uses at least.
Jens: mmhm
Nina: both the English and
Jens: mhm
Nina: and it is true about the language. It is not.
Jens: Yes, but, you are right. Jeg have probably focused more ehrm on the language. Nina: yes, yes
Jens: compared to the content her, because because these are. There are a number of instances where pronouns are used without it being clear what is referred to. It really annoys me.
(Nina and Jens, script 5)

Others voiced their insecurities and thereby sought clarification or help:

Jens: det jeg har, eller jeg har været usikker. Usikkerheden er gået på at det er at jeg kan se at jeg har skrevet en del hvor sådan jeg synes sådan der er sammenhæng i de forskellige tekst. Atså afsnittene er for så vidt indholdsmæssigt hver for sig
Nina: hmm
Jens: altså meget, rimelig fornuftige.
Nina: hmmm
Jens: ehrm og der er også rel, relativ få fejl.
Nina: hmmm
Jens: Jeg synes ikke der, der er den der indre sammenhæng i meningen. Den går fra første til andet afsnit. Fra andet til tredie jeg synes ikke at. Det er ligesom det er løsrevne, sådan ehrm, sådan meninger.
(Nina and Jens, script 12)

Translation:
Jens: here I have, or here I have been unsure. I was unsure because I can see that I have written down notes about what I think, think there is coherence in the different texts. I mean the paragraphs are in themselves content.
Nina: hmm
Jens: they really make a lot of sense
Nina: hmmm
Jens: ehrm and there are also rel, relatively few errors.
Nina: hmmm
Jens: I don't think there, there is this inner coherence. It is ok from the first paragraph to the second one. From the second one to the third one I don't think that. It is as if they are detached, the, ehrm, the opinions
(Nina and Jens, script 12)

The raters' deliberation of their own assessment strategies was occasionally accompanied by reference to the raters' own teaching practices. In this way, they also potentially developed their

own teaching skills or awareness about them (Deliberate/Articulate Teaching Strategies/Practices: *M*=0.7%):

> *Malene*: altså det, altså det er i det hele taget be besynderligt, for vi får alle mulige andre historier end dem som opgavesættet handler om
> *Thea*: ja
> *Malene*: i virkeligheden
> *Thea*: ja
> *Malene*: og der ville jeg så sige til mine egne elever at at det det er helt fint hvis de kan supplere med noget andet
> *Thea*: ja
> *Malene*: men de skal
> *Thea*: ja, men de skal inddrage. Og der har jeg måske nok så mere tænkt, nå men der er trods alt blevet skrevet en hel del.
> (Malene and Thea, script 15)
>
> *Translation:*
> *Malene: it really is, it really is rather odd, for we get all kinds of other stories than the ones the source materials are about.*
> *Thea: yes*
> *Malene: as a matter of fact*
> *Thea: yes*
> *Malene: and in such cases I would say to my own students that it is perfectly fine to include something else*
> *Thea: yes*
> *Malene: but they have to*
> *Thea: yes, but they must include. And in this case I probably thought to myself, well but at least they have written quite a bit.*
> *(Malene and Thea, script 15)*

5.2.2.3 Justifying Judgments: Textual Focus

In adopting a textual focus during the communal ratings, the raters balanced their attention among a variety of different text features.

5.2.2.3.1 Judging Content and Use of Source Materials

Of the textual features, Content and Use of Source Materials attracted the most attention. An average of 19.6% of the decision-making behaviors focused on judging the Content and Use of Source Materials. More specifically the raters would focus on the students' line of reasoning (Assess or Justify Reasoning, Logic, or Topic Development: *M*=3.6%), for example:

Jette: ja, netop. Og indholdet er også meget selvstændigt, synes jeg. Den er logisk bygget op. Den har nogle rigtig gode argumenter. Måske kommer man for hurtigt ind på det der med 9/11.
Ken: ja, jeg forstår godt hvad du mener. Det er da også en spændende teaser i begyndelsen
(Jette and Ken, script 6)

Translation:
Jette: Yes, exactly. And the content is also more independent, I think. It has a logical line of thought. It has some really good arguments. Perhaps they present the issue about 9/11 too early.
Ken: Yes, I know what you mean. There is also an exciting teaser in the beginning.
(Jette and Ken, script 6)

Or they focused on the relevance of information expressed in the compositions (Assess or Justify Task/Topic Relevance or Completion: M=3.5%)

Julie: Der er noget underligt i afsnit nummer tre, ehrm.
Gitte: ja, jeg
Julie: der bliver inddraget nogle ting omkring 9/11.
Gitte: ja,
Julie: som vi ikke rigtig
Gitte: nej
Julie: ja, det afviger
Gitte: nej, det afsnit har jeg netop sat kæmpe spørgsmålstegn ved og sagt "er det relevant?".
Julie: ja
Gitte: ehrm, ehrm, lige pludselig man taler om børnearbejde, som vedkommende gør i afsnit 2 og sådan noget, ikke? Der er det koncentreret, centreret omkring børnearbejde, så kommer der så noget, et afsnit her om 9/11. ehrm og noget med muslimer og sådan noget. Der føler jeg er sådan lidt surt opstød på en eller anden led. Det er ikke relevant.
(Gitte and Julie, script 3)

Translation:
Julie: there is something odd in paragraph three, ehrm.
Gitte: yes, I
Julie: some things about 9/11 are being included.
Gitte: yes.
Julie: that we don't really
Gitte: no
Julie: yes, is different from
Gitte: no, I have put a big question mark next to this paragraph and said "it this relevant?"
Julie: yes
Gitte: ehrm, ehrm, all of a sudden they talk about child labor, as the test taker does in paragraph two and things like that, right? Here they concentrate on, focus on child labor, then comes something, a paragraph 9/11. ehrm and something about Muslims and things like that. Here I feel offended in a way. It is no relevant.
(Gitte and Julie, script 3)

In addition, and relatedly, the raters focused on the student's uses of the source materials (Assess or Justify Use and Understanding of Source Material: *M*=3.0%):

> *Louise*: AND BY CHILDREN WHO ARE OLD ENOUGH. Fordi det er jo netop det der ikke. Altså der står jo at de har jo ikke brugt det i deres markedsføring.
> *Lone*: præcis, det er jo en tekstmisforståelse, ikke?
> *Louise*: ja.
> (Louise and Lone, script 14)
>
> *Translation:*
> *Louise: AND BY CHILDREN WHO ARE OLD ENOUGH. Because that is really it, isn't it? It actually says that they did not use it in their marketing campaign.*
> *Lone: precisely, it is, in fact, a misunderstanding of the source texts, isn't it?*
> *Louise: yes*
> *(Louise and Lone, script 14)*

Impressions of the student's independence or maturity level were also brought into the discussion (Assess or Justify Task/Topic Maturity or Independence: *M*=2.6%), at times providing the final reason for the score:

> *Ken*: Ja, og det gør hun på en fornuftig måde. På en selvstændig måde. Jeg kan også godt lide at personen sådan stiller spørgsmål ind i mellem for at gøre det lidt mere levende.
> *Jette*: ja, det er rigtigt, men vi kan da godt give den 10 for min skyld. For, som du siger, den behandler emnet selvstændigt. Altså.
> *Ken*: ja, det synes jeg vi skal. Ok.
> *Jette*: ja, fint 10 til opgave 4.
> (Jette and Ken, script 4)
>
> *Translation:*
> *Ken: yes, and she does it in a way that makes sense. In an independent way. I also like the fact that the person like asks questions here and there to make it a bit more alive.*
> *Jette: yes, it is true, but I don't' mind giving it a 10. For as you say, it treats the topic independently. So ehrm.*
> *Ken: yes, I think we should do that, ok.*
> *Jette: yes, a nice 10 to script 4*
> *(Jette and Ken, script 4)*

Likewise, the raters judged the students writers' clarity (Assess or Justify Clarity: *M*=2.4%) and correctness of content (Assess or Justify Correctness of or Disagreement of Content: *M*=1.0%):

> *Gitte*: nej, der er i hvert fald noget tekst. Vedkommende har forsøgt
> *Julie*: men altså, jeg har skrevet uklart budskab, forvrøvlet indhold og bruge tekstoplysninger til noget.

Gitte: nej
Julie: og forkerte eller selvlavede oplysninger.
...
Gitte: ja, jeg har også godt nok lavet en lille 5'er. Så om ehrm, for der står faktisk, det er også noget vrøvl, der står i den.
Julie: ja, det er det.
(Gitte and Julie, script 11)

Translation:
Gitte: no, at least there is some text. The test taker has tried.
Julie: but really, I have written unclear message, muddled content and use of the information from the source texts for something.
Gitte: no
Julie: and uses wrong or self-constructed information.
...
Gitte: yes, jeg have also written a low 5. So ehrm, for it actually says, it is also nonsense what's in it.
Julie: yes, it is
(Gitte and Julie, script 11)

5.2.2.3.2 Judging Language

Slightly less focus was put on textual features related to language. About 15.7% of the raters' decision making focused on judging the language of the script. Besides judging the language overall, attention was paid mostly to errors (Assess or Justify Frequency of Errors: M=3.6%; Assess or Justify Gravity of Errors: M=1.1):

Jens: det har jeg også skrevet, det er. Men det er selvfølgelig også med i lyset af det foregående, ikke. At der er relativt få fejl. Men det er der vel reelt. Altså, der er selvfølgelig nogle kongruensfejl. Det er der snart i alle opgaver.
Nina: ja, ja
Jens: men ellers så er det jo meget få og ikke sådan særlig tunge fejl.
Nina: nej, nej
(Nina and Jens, script 4)

Translation:
Jens: I wrote that too, it is. But it is, of course, also in light of the previous ones, right. That there are relatively few errors. But in fact, there is. I mean, of course some subject-verb agreement errors. We find that nearly in all essays now.
Nina: yes, yes
Jens: but otherwise there are few and not very serious, grave errors.
Nina: no, no
(Nina and Jens, script 4)

Language errors that were mentioned were often subject-verb agreement (as illustrated in Nina's and Jens' interaction above) or syntax:

> *Hans*: THE EMBARRASSING SUBJECT THAT IT HAS BECOME
> *Henrik*: ja
> *Hans*: der synes jeg at sætningsopbygningen er meget simpel og noget der lyder dansk, og
> *Henrik*: ja, en lille smule danisme.
> (Hans and Henrik, script 4)
>
> *Translation:*
> *Hans: THE EMBARRASSING SUBJECT THAT IT HAS BECOME*
> *Henrik: yes*
> *Hans: here I simply think that the syntax is very simple, and something that sounds Danish, and*
> *Henrik: yes, a bit of danishm.*
> *(Hans and Henrik, script 4)*

Attention to the comprehensibility of the language was also one of the favored language-related foci (Assess or Justify Fluency or Comprehensibility: M=3.1%):

> *Susanne*: jeg synes altså heller ikke at sproget kan bære mere end. Der er nogle sætninger som er meget svært forståelige.
> *Tove*: ja, der er der nemlig. Det er uklare formuleringer.
> *Susanne*: meget uklare formuleringer
> *Tove*: og CHILDREN WORK og har udviddet tid, hvor det ikke hører hjemme.
> *Susanne*: ja
> *Tove*: jeg synes ikke den kan bære mere end et 7-tal.
> *Susanne*: Ja, det bliver et 7-tal.
> (Susanne and Tove, script 3)
>
> *Translation:*
> *Susanne: I really don't think that the language is more than. Some of the sentences are very difficult to understand.*
> *Tove: yes, exactly. Unclear wordings.*
> *Susanne: very unclear phrases..*
> *Tove: and CHILDREN WORK and uses the progressive tense where it is not supposed to be.*
> *Susanne: yes*
> *Tove: I don't think we can give it more than a 7.*
> *Susanne: Yes, it is a 7. .*
> *(Susanne and Tove, script 3)*

5.2.2.3.3 Judging Organizational Structure, Style/Format, and Amount of Text

Although less pronounced than the focus on Language and Content, attention was also paid to Organizational Structure (an average of 7.7% of decision making behavior was devoted to judging the Organizational Structure of the scripts) and to Style and Format (M=4.1%):

> *Ken*: og det jeg godt kan lide ved den er at den er bygget rigtig godt op, med god indledning, som passer til overskriften. Og overskriften tages op igen.
> *Jette*: ja, det gør den. Ja, og man kan se at personen kan finde ud af at lave en sammenhængende opgave.
> *Ken*: ja, meget sammenhængende.
> *Jette*: eleven kan skrive så det passer ind i script genren.
> (Jette and Ken, script 4)
>
> *Translation:*
> *Ken: and I like that it has a good structure, with the introduction that fits the title. And the title is taken up again further down.*
> *Jette: yes, it is. Yes, and it is obvious that the test taker knows how to construct a coherent text.*
> *Ken: yes, very coherent.*
> *Jette: the student can write in this genre.*
> *(Jette and Ken, script 4)*

Attention was given as well as to Amount of Text (M=4.1%). As with the independent ratings, during the communal ratings, Amount of Text was at times cited as the main reason for giving a final score:

> *Thea*: der er sådan lidt mere nuance på trods alt, ikke?
> *Malene*: ja,
> *Thea*: så derfor synes jeg det, men den er så ikke meget fyldig, kan man sige.
> *Malene*: nej, nej,
> *Thea*: og derfor så kan den aldrig komme højere op.
> *Malene*: nej, men det er jo det så.
> (Malene and Thea, script, 13)
>
> *Translation:*
> *Thea: It is after all a bit more advanced, right?*
> *Malene: yes*
> *Thea: so that is why I think so, but you could say that it is not very long.*
> *Malene: no, no*
> *Thea: and that is why it can never be a higher score.*
> *Malene: no, but this is so, isn't it.*
> *(Malene and Thea, script, 13)*

5.2.2.4 Summary of Raters' Judgment Behaviors in Communal Ratings

In judging their images of the student scripts together in dyads the raters expressed a variety of judgment strategies focusing on a mixture of textual and contextual features as well as rater monitoring aspects. Spending considerable energy to articulate scores, the raters focused on their common task of assigning scores. However, discussions about score assignment did not appear to be arguments or disputes. Rather, in cases of score discrepancy the raters would go through an elaborate justification process in which they validated their personal perceptions of the scripts by balancing their focus among the different textual features and supporting their claims by exemplifying directly from the student scripts, comparing them, or giving their overall impressions. If, after deliberating and justifying their assessments, the raters did not reach a consensus, the scores were not finalized by consulting the scale descriptors directly. Rather, the score was resolved by mutual compromise. Potentials for rater development also appeared in these discussions as the raters voiced their insecurities, defined, suggested or revised their assessment strategies, and at times even referred to their own teaching practices.

5.3 Sequence of Decision-Making Behaviors

As with the independent rating sessions, a typical sequence of decision-making seemed to appear in the communal rating sessions. The phases in the communal rating sessions were not as clearly demarcated as the three phases were in the independent rating sessions, but the following three phases were evident:

Phase 1: Exchange preliminary scores from independent rating sessions. This phase was used by all raters and with all scripts. The phase was very brief and consisted of the raters exchanging their original independent scores.

Phase 2: Deliberate and justify scores. Here the raters would exchange assessments of the student scripts, occasionally seeking clarification from their co-rater, deliberate assessment strategies, and bringing hard evidence to the table by exemplifying directly from the scripts. Focus was on language, content, but also organizational structure, style/format and amount of text. Rather than consulting the scale descriptors directly, the raters looked to their co-raters for

help when in doubt. This phase was used by all raters and always in cases of discrepancy in the raters' original independent scores. It was also sometimes used when the raters agreed, although the phase tended to be shorter than if the raters disagreed.

Phase 3: Finalize scoring. The final phase was used by all raters and with all scripts. The phase was very brief and focused on reaching a final score. In case of full agreement on the original independent scores, the final score remained the same as the independent scores. In case of adjacent scores, sometimes the higher score was chosen and sometimes the lower score was chosen. This decision relied on the justification phase. In the rare case of a two-point discrepancy in the original independent scores, the final score selected was most often the score midway between the independent scores, although at times the raters decided that the most appropriate score would be one of the independent scores, or even lower than the lower of the independent scores.

These three phases were enacted by all rater dyads, although Phase 2 at times was quickly glossed over or even skipped in cases of full agreement between the co-raters. Below is a prototypical example of a communal rating session:

Phase 1: *Jette*: opgave 6. Den synes jeg var god.
Ken: ja
Jette: jeg har givet den et lille 10-tal, altså
Ken: ok, der ligger jeg så et tak under dig for jeg har et lille 9-tal.

Phase 2: Jeg kan godt se at det er en god opgave, men den er altså for kort.
Jette: jo, det er den desværre.
Ken: ja, ærgeligt, men den er ikke mere end en side, og det synes jeg altså burde trække ned.
Jette: joh, ja, du har ret. Men jeg synes den har nogle ret flotte passager, hvor sproget er meget elegant. Det flyder og nogle meget fornemme ordvalg.
Ken: jo, det er rigtigt. EMBARKING ON A CRUSADE
Jette: ja, netop. Og indholdet er også meget selvstændigt, synes jeg. Den er logisk bygget op. den har nogle rigtig gode argumenter. Måske kommer man for hurtigt ind på det der med 9/11.
Ken: ja, jeg forstår godt hvad du mener. Det er da også en spændende teaser i begyndelsen. Det er sådan meget metaforisk sprog. Det kan jeg godt se.
Jette: men jo du har ret, den er nok for kort til

Phase, 3: *Ken*: ja, til at vi skal helt op på 10. Jo, jeg synes vi skal give den 9.
Jette: ja, det synes jeg passer bedre. 9 til 6'eren.
(Jette and Ken, script 6)

Translation:

Phase 1: Jette: script 6. I think this was good.
Ken: yes
Jette: well, I have given it a low 10.
Ken: ok, I am a level below you because I have given it a 9.

Phase 2: I see that it is a good essay, but it is, in fact, too short.
Jette: yes it is, unfortunately.
Ken: yes, too bad, but it is not more than a page, and I really think that points should be taken off for that.
Jette: yeah, yes, you are right. But I think it has some really beautiful passages where the language is very elegant. It is fluent and very elegant choice of words.
Ken: yes, it is true. EMBARKING ON A CRUSADE
Jette: yes, exactly. And the content also more independent, I think. It has a logic line of thought. It has some really good arguments. Perhaps they present the issue about 9/11 too early.
Ken: yes, I know what you mean. There is also an exciting teaser in the beginning.
It is well a very metaphorical language. I see that
Jette: but you are right, it is probably too short for

Phase 3:Ken: yes, for going all the way up to a 10. Yes, I think we should give it a 9.
Jette: yes, I think that is more appropriate. A 9 for number 6.
(Jette and Ken, script 6)

5.3.1 Summary of the Sequence of Raters' Decision-Making Behaviors in Communal Ratings

The sequence of rating behavior during the communal ratings is synthesized in Figure 5.1.

5.4 Balance of Attention to Official Assessment Criteria

As with the analysis of the decision-making behaviors in the independent ratings, the coding scheme facilitated insights into how the raters distributed their attention to textual features corresponding to the official assessment criteria in the communal ratings. Figure 5.2 illustrates the raters' distribution of attention to these features.

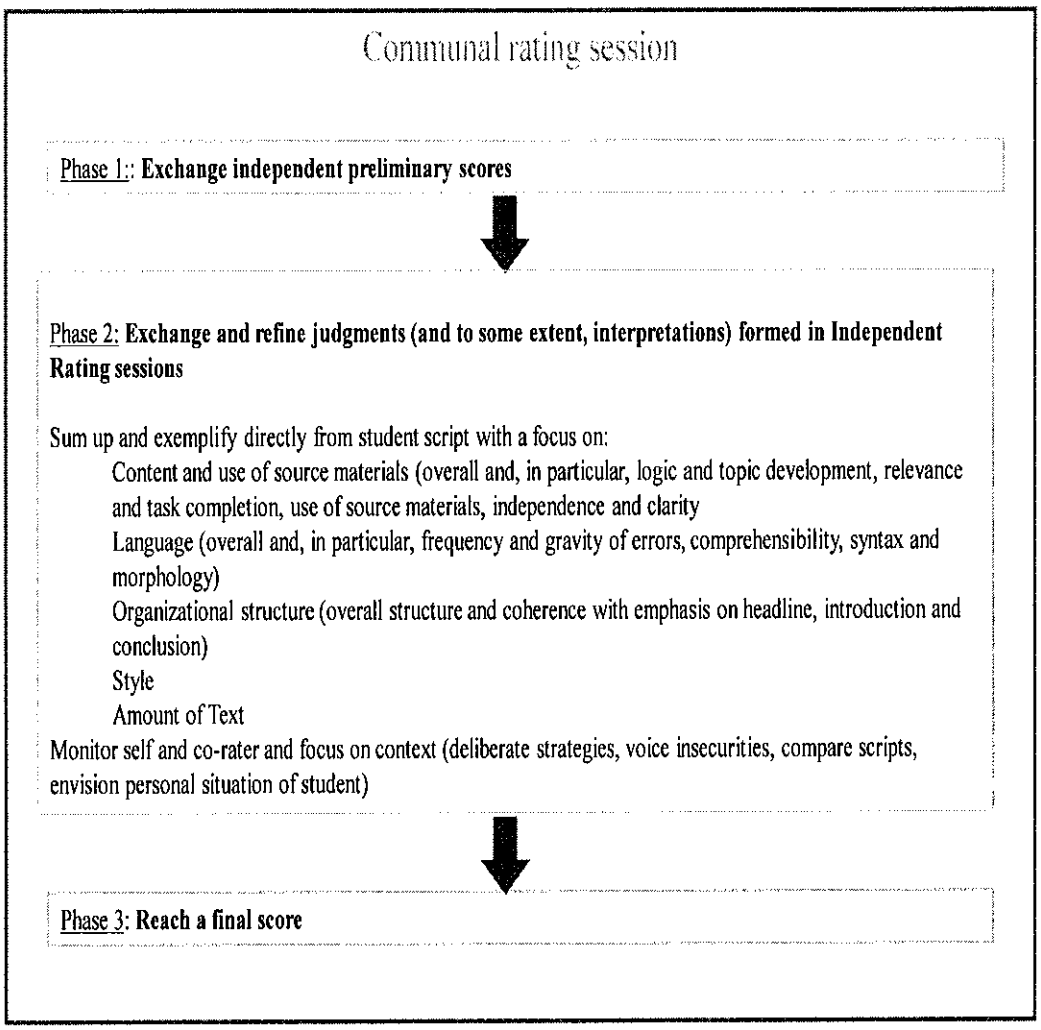


Figure 5.1: Prototypical Sequence of Raters' Decision-Making Behaviors in Communal Rating Sessions

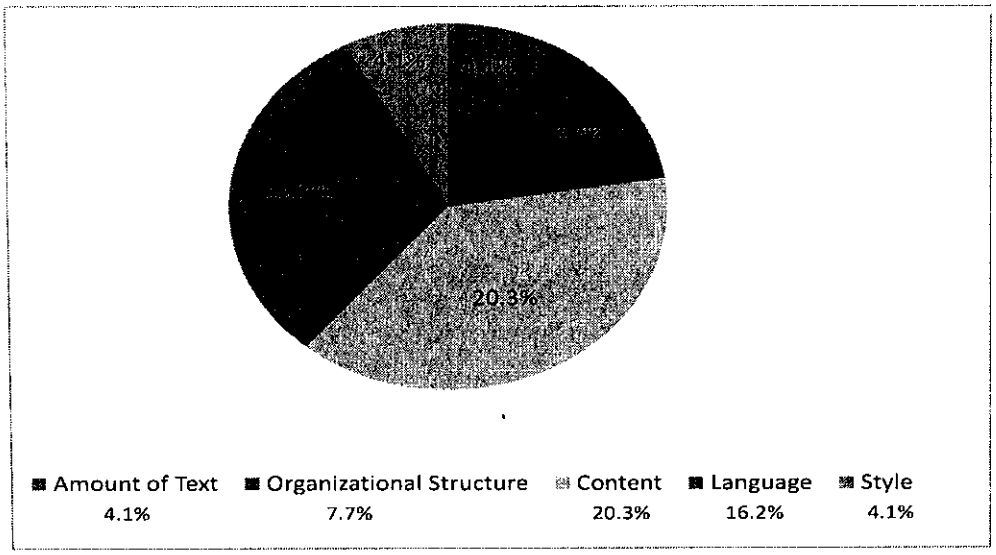


Figure 5.2: Mean Percentages of Textual Features Attended to in Communal Ratings

As can be seen from Figure 5.2, of the textual features mentioned during discussions in communal ratings, Content and Use of Source Materials attracted the most attention (M=20.3%), followed by Language (M=16.2%). The raters also attended to Organizational Structure (M=7.7%) and less to Style and Format (M=4.1%) and Amount of Text (M=4.1%). Thus the distribution of attention to textual features corresponding to the official assessment criteria for the exam seemed to be relatively well-balanced.

5.5 Equality of Engagement

To determine the extent to which the raters in this study engaged equally in the communal rater discussions, levels of score dominance and conversational dominance were reported.

5.5.1 Score Dominance

As described in Chapter 3, score dominance refers to the distance between the independent scores of each of the two raters and the final, communally rated score. Score dominance is here presented by the extent to which the final score was closer to the independent score of one rater (score dominating rater) or the other (score conceding rater) in the 10 rating dyads.

A total of 150 final scores were assigned (10 rater dyads rating 15 student scripts each). In 70 of these cases the raters came to the communal rating session dyads with similar independent scores, and so, of course, the independent scores became the final score. In 80 of the cases, however, the raters entered the dyads with discrepant independent scores. In 15 of these discrepant score cases, the final communal score became a compromise between the independent scores, i.e. the final score fell midway between the two raters' independent scores. In 65 of these discrepant score cases, however, one rater conceded his/her score, i.e. the final score was further away from the independent score of one rater than the independent score of the other rater. The relatively high number of concession cases is a product of the adjacency of scores: in most cases one of the raters had to concede to the other because their original independent scores were adjacent (e.g. one rater had an 8 as the original independent score, and the other a 9).

Table 5.4 shows the score dominating/conceding behaviors of each of the raters. The first two columns represent the rater dyad and the raters. The third column, Number of Score Dominations, indicates the number of times each rater dominated by score in his/her rater dyad (i.e., each time his/her original independent score was closer to the final score), and the corresponding Number of Score Concessions indicates the number of times each rater conceded his/her score (i.e., each time his/her original independent score was further away from the final score). The fifth column shows the number of compromises made between the raters in each of the dyads, the sixth column sums up the number of score discrepancies made in the dyads, and the last column totals the number of scores assigned.

Table 5.4: Score Dominance in Communal Ratings

Rater Dyad	Rater	Number of Score Dominations	Number of Score Concessions	Number of Compromises	Total Number of Discrepancies	Total Number of Final Scores Assigned
1	**Pernille**	3	7	2	12	15
	Jesper	7	3			
2	**Gitte**	6	3	0	9	15
	Julie	3	6			
3	**Torben**	2	1	3	6	15
	Tina	1	2			
4	**Tove**	0	1	1	2	15
	Susanne	1	0			
5	**Nina**	2	3	2	7	15
	Jens	3	2			
6	**Lone**	3	1	4	8	15
	Louise	1	3			
7	**Astrid**	9	1	0	10	15
	Helle	1	9			
8	**Thea**	6	3	0	9	15
	Malene	3	6			
9	**Jette**	3	5	0	8	15
	Ken	5	3			
10	**Hans**	2	4	3	9	15
	Henrik	4	2			
	Total	*65*	*65*	*15*	*80*	*150*

Table 5.4 shows that some raters did tend to exhibit more score dominance than others. Using a definition of a score dominator as someone whose score dominates in more than half of the concession cases, there were 3 score dominators out of a total of 20 raters, although none of these raters dominated by score in all the concession cases. The score dominators were:

- Jesper, who dominated seven times over his co-rater Pernille, who dominated over him only three times,
- Lone, who dominated three times over her co-rater Louise, who dominated over her only once, and

- Astrid, who dominated nine times over her co-rater Helle, who dominated over her only once.

5.5.2 Conversational Dominance

Conversational Dominance refers to the extent to which each rater participates co-equally in the rater conversations, here operationalized as number of words and decision-making behaviors made by each rater in the rater dyads (see Chapter 3).

Table 5.5 summarizes the rater dyads' conversational dominance operationalized as the number of words. Table 5.6 displays their conversational dominance operationalized as the number of decision-making behaviors. The third column in both Tables shows the share (in percent) each rater in the rater dyads had of the overall amount of words and decision-making behaviors respectively. The fourth column in both Tables shows the difference between the raters' share of number of words and number of decision-making behaviors respectively.

As can be seen from Table 5.5, there is an average difference of 15.9 percent between the two raters in the rater dyads with respect to the amount of words spoken in the communal rating sessions. The differences range from 5.4 percent to 38.6 percent. This pattern is echoed in the distribution of decision-making behaviors (Table 5.6), the average difference between the raters in the rater dyads being 1.63, ranging from 3.0 percent to 30.9 percent. These numbers do not suggest much of a difference in the number of words or number of decision-making behaviors between two raters in the rater dyads. This appears to indicate that one rater did not dominate the communal rating sessions completely. Ironically, the rater who seemed to dominate the conversation the most (Hans) tended to concede his scores.

Table 5.5: Number of Words in Communal Ratings (spoken per script by each rater in the 10 rater dyads)

Rater Dyad	Number of Words, *M* *(SD)*	Percent of Total Amount of Words	Difference in Number of Words (percent)
1. Pernille	199.1 (162.1)	41.9	16.3
1. Jesper	188.9 (128.6)	58.1	
2. Gitte	187.1 (81.3)	61.9	23.8
2. Julie	115.1 (77.5)	38.1	
3. Torben	258.2 (162.8)	46.9	6.2
3. Tina	292.1 (150.4)	53.1	
4. Tove	195.2 (59.5)	58.1	16.3
4. Susanne	140.5 (54.1)	41.9	
5. Nina	204.3 (134.5)	62.3	24.5
5. Jens	336.9 (211.1)	37.7	
6. Lone	108.7 (56.4)	44.7	10.5
6. Louise	88.0 (51.2)	55.3	
7. Astrid	284.6 (261.7)	55.9	11.7
7. Helle	224.8 (161.4)	44.1	
8. Thea	163.1 (115.7)	40.5	19.0
8. Malene	239.7 (141.0)	59.5	
9. Jette	93.0 (43.4)	52.7	5.4
9. Ken	83.4 (29.1)	47.3	
10. Hans	232.6 (91.5)	69.3	38.6
10. Henrik	103.1 (38.6)	30.7	
Mean(SD)			15.9 (11.)

Table 5.6: Number of Decision-Making Behaviors in Communal Ratings (produced per script by each rater in the 10 rater dyads)

Rater Dyad	Number of Decision-Making Behaviors, *M (SD)*	Percent of Total Amount of Decision-Making Behaviors	Difference in Number of Decision-Making Behaviors (percent)
1. Pernille	9.2 (6.2)	52.5	4.9
1. Jesper	8.3 (5.1)	47.5	
2. Gitte	11.3 (3.7)	59.9	19.9
2. Julie	7.5 (3.4)	40.1	
3. Torben	11.6 (4.8)	45.3	9.4
3. Tina	14.0 (4.8)	54.7	
4. Tove	11.1 (4.6)	58.2	16.5
4. Susanne	7.9 (2.8)	41.8	
5. Nina	10.4 (6.0)	41.6	16.8
5. Jens	14.6 (6.2)	58.4	
6. Lone	7.3 (3.2)	59.2	18.5
6. Louise	5.0 (2.6)	40.8	
7. Astrid	16.3 (9.6)	60.2	20.5
7. Helle	10.7 (4.3)	39.8	
8. Thea	7.4 (4.0)	38.9	22.1
8. Malene	11.6 (5.8)	61.1	
9. Jette	6.9 (3.3)	51.5	3.0
9. Ken	6.5 (1.4)	48.5	
10. Hans	13.3 (3.2)	65.5	30.9
10. Henrik	7.0 (2.9)	34.5	
Mean(SD)			16.3 (8.0)

5.5.3 The Relationship between Score Dominance and Conversational Dominance

To determine whether the rater who dominated by score also dominated the conversation within the dyads, I examined each case of score dominance for conversational dominance. Table 5.7 below shows the relationship between score dominance and conversational dominance.

Table 5.7: Mean (M) Number of Words and Decision-Making Behaviors and Standard Deviations (SD) in Score Dominating Cases and in Score Conceding Cases

	Verbosity (words per script)	Number of decision-making behaviors (per script)
Score dominating cases n = 65	204.6 (154.6)	9.9 (4.8)
Score conceding cases n = 65	216.9 (132.6)	11.2 (5.4)

As can be seen from Table 5.7, there was very little difference in verbosity and number of decision-making behaviors between the score dominating and score conceding cases. If a rater dominated a script by score (i.e., if his/her independent score came closer to the final, communal score), he/she would produce an average of 204.6 words and 9.9 decision-making behaviors during the rating process. If, on the other hand, a rater conceded his/her score (i.e., if his/her independent score was further away from the final, communal score), he/she would produce an average of 216.9 words and 11.2 decision-making behaviors during the rating process. This small difference in verbosity and number of decision-making behaviors in score dominating and score conceding cases implies that score dominance was not a product of conversational dominance, and indicates that the communal rating sessions in this study were characterized by equal engagement between the raters in the dyads.

5.5.4 Summary of Equality of Engagement

Although some of the raters tended to dominate their dyad by score, score dominance seemed not to be a product of conversational dominance. The raters in the dyads appeared to be equally engaged in their rating conversations as indicated by the number of words and number of decision-making behaviors. Further, and perhaps more importantly, in cases where one rater dominated by score, that person did not dominate the conversation within the dyad.

5.6 Rater Agreement

I calculated the level of agreement among the rater dyads using the same formulas as in the independent ratings. Kendall's *W* test for the communal ratings is shown in Table 5.8 below.

Table 5.8: Kendall's W in Communal Ratings

N	10
Kendall's *W*	.90
Chi-Square	126.249
Df	14
Asymp. Sig.	.000

As show in Table 5.8, Kendall's *W* showed an agreement of .90 for 150 scoring decisions (1 indicates full agreement and 0 indicates no agreement at all). Cronbach's Alpha showed an agreement of $\alpha = .80$. As with the independent ratings, for the communal ratings, Kendall's showed a higher agreement rate than Cronbach's Alpha. This suggests that the rater dyads in this study were in high agreement in terms of their rank-ordering the scripts. Nonetheless, they exhibited some degree of variance in the scores they assigned to the scripts.

5.7 Summary of Raters' Decision-Making Behaviors in Communal Ratings

Although the raters spent time revisiting the student scripts to confirm or refute their initial, individual interpretations, primarily by deciphering unclear phrases, the communal rating sessions were characterized by a process of justifying and deliberating judgments of the scripts and finalizing their scores. The raters would start the communal rating sessions by exchanging with each other their preliminary scores from their independent rating sessions. They then continued into a justification phase, particularly in cases of score discrepancy. Here the raters validated their judgments, balancing their attention to different types of textual criteria corresponding to the official assessment criteria. They would supplement their textual focus by attending to contextual factors such as envisioning the personal situation of the student. To justify their claims they would often exemplify directly from the student scripts, compare scripts,

and give their general impressions. Insecurities were expressed and discussed as were general assessment strategies. However, they never attempted to resolve their judgments or insecurities by consulting the scale descriptors in the scoring rubric, but rather relied on each other for resolution. The final score was assigned in the final phase of communal rating. This was sometimes the result of a consensus and sometimes of a compromise.

The raters balanced their attention relatively evenly to the textual features corresponding to the official assessment criteria, even placing slightly more focus on content and use of source materials than on language.

In resolving their differences the raters seemed to engage equally in the rater discussions. Although three of the raters dominated by score, score dominance did not appear to be a product of conversational dominance as the raters whose score dominated in the particular score resolution cases did not produce more words or more decision-making behaviors than did the raters who conceded their scores.

With respect to rater agreement levels, the communally rated scores produced an agreement level of .90 (Kendall's W) and .80 (Cronbach's Alpha).

Chapter 6

Chronicling Raters' Decision-Making Behaviors from their Independent Rating Sessions to their Communal Rating Sessions

6.1 Purpose and Scope of the Chapter

This chapter chronicles the progression of the raters' decision-making, tracing and comparing their behaviors from their independent rating sessions to their communal rating sessions. In doing this the findings from Chapter 4: Raters' Decision-Making Behaviors in Independent Rating Sessions and Chapter 5: Raters' Decision-Making Behaviors in Communal Rating Sessions are reproduced briefly and compared here. The progression and comparison of the two rating sessions is introduced by presenting the trends in the raters' distinct decision-making behaviors in the two sessions, including the sequence of these behaviors. A comparison is also made of how the raters balanced their attention to the textual features corresponding to the official assessment criteria. Next the raters' score ranges and score agreement levels in the two sessions are compared, and finally the raters' own perceptions of the score progression in the two rating sessions are reported along with their general perceptions of CWA.

6.2 Trends in Raters' Distinct Decision-Making Behaviors from the Independent Ratings to the Communal Ratings

Most of the decision-making behaviors displayed in the independent rating sessions were repeated in the communal rating sessions (see Table 6.1, which combines Tables 4.1 and 5.1, the results from the independent ratings and the communal ratings respectively), although the distribution of these behaviors varied from one session to the other.

Table 6.1: Grand Mean Percentages (M) and Standard Deviations (SD) of Raters' Decision-Making Behaviors Rating the Same Scripts first in Independent Rating Session, then in Communal Rating Sessions

	Independent Ratings: 300 protocols (20 raters rating 15 scripts)	***Communal Ratings: 300 protocols (20 raters rating 15 scripts)***
	M(SD)	***M(SD)***
Interpretation Strategies		
Contextual or Monitoring Focus		
Read or interpret task input/source materials	0.1% (0.4)	0.2% (0.8)
Read or reread student script	29.3% (12.2)	0.0% (0.0)
Envision personal situation of the student	0.8% (1.5)	2.9% (6.2)
Consider task or exam requirements	0.2% (0.6)	1.2% (3.1)
Consider own perception of correct English (e.g. consult a dictionary)	0.2% (0.8)	0.1% (0.4)
Textual Focus (Amount of text)		
Scan script for length	0.2% (0.7)	0.0% (0.0)
Textual Focus (Organizational Structure)		
Discern or scan for organizational structure	1.2% (2.0)	0.0% (0.0)
Textual Focus (Content and Use of Source Materials)		
Discern or summarize ideas	1.8% (2.6)	0.1% (0.5)
Identify or interpret ambiguous or unclear phrases	2.8% (3.1)	0.6% (1.6)
Textual Focus (Language)		
Classify language errors into types	7.9% (7.8)	0.0% (0.0)
Identify errors	4.8% (5.2)	0.1% (0.2)
Correct or edit language (errors or unclear phrases)	13.4% (9.6)	0.4% (1.1)
Textual Focus (Style and Format		
Discern style, register or genre	0.7% (1.5)	0.1% (0.2%)
Judgment Strategies		
Contextual or Monitoring Focus		
Articulate score	3.1% (2.3)	21.7% (12.5)
Compare student script	0.7% (1.5)	2.8% (5.2)
Define, revise or suggest assessment strategies	0.4%(1.5)	6,6% (8.3)
Articulate general impression	1.8% (2.8)	3.3% (5.4)
Deliberate/articulate teaching strategies/practices	0.1% (0.4)	0.7% (2.2)
Exemplify directly from student script	1.4% (5.3)	7.4% (8.7)
Consider consensus-based strategy	0.0% (0.3)	0.5% (1.4)
Consider personal response or bias	0.0% (0.0)	0.3% (1.0)

Textual Focus (Amount of Text)		
Assess or justify amount of text	1.8% (2.3)	4.1% (6.5)
Textual Focus (Organizational Structure)		
Assess or justify organizational structure overall	0.7% (1.4)	1.4% (3.2)
Assess or justify title	1.2% (1.7)	2.3% (5.0)
Assess or justify introduction and/or conclusion	1.3% (1.8)	1.8% (4.0)
Assess or justify coherence and/or cohesion	0.8% (1.6)	2.1% (4.0)
Textual Focus (Content and Use of Source Materials)		
Assess or justify content/ideas overall	1.2% (1.9)	3.5% (5.8)
Assess or justify reasoning, logic, or topic development	1.6% (2.4)	3.6% (6.2)
Assess or justify clarity	0.3% (0.8)	2.4% (4.4)
Assess or justify correctness of or disagreement with content	1.5% (2.5)	1.0% (2.3)
Assess or justify maturity or independence	0.9% (1.8)	2.6% (4.9)
Assess or justify task/topic relevance or completion	2.0% (2.9)	3.5% (5.8)
Assess or justify use and understanding of source material	1.7% (2.6)	3.0% (5.3)
Textual Focus (Language)		
Assess or justify language overall	2.6% (2.9)	4.0% (6.6)
Assess or justify frequency of errors	2.1% (2.4)	3.6% (6.2)
Assess or justify gravity of errors	0.8% (1.6)	1.1% (2.6)
Assess or justify syntax or morphology	2.5% (3.6)	2.1% (4.6)
Assess or justify lexis	1.4% (2.4)	0.6% (2.1)
Assess or justify fluency or comprehensibility	3.4% (3.2)	3.1% (6.1)
Assess or justify spelling	0.4% (1.2)	1.1% (2.8)
Assess or justify punctuation	0.3% (0.9)	0.0% (0.0)
Textual Focus (Style and Format)		
Assess or justify style	0.5% (1.3)	2.5% (5.7)
Assess or justify genre	0.4% (1.2)	1.6% (3.8)

Table 6.2: Grand Mean Percentages (M) and Standard Deviations (SD) of Raters' Decision-Making Behaviors Rating the Same Scripts first in Independent Rating Sessions (20 raters, 15 scripts=300 protocols), then in Communal Rating Sessions (20 raters, 15 scripts=300 protocols)

	Monitoring Focus		**Textual Focus**										**Total**	
			Amount of Text		**Organizational Structure**		**Content and Use of Source Material**		**Language**		**Style and Format**			
	Ind.	**Com.**	**Ind.**	**Com.**	**Ind.**	**Com.**	**Ind.**	**Com.**	**Ind.**	**Com.**	**Ind.**	**Com.**	**Ind.**	**Com.**
Interpretation Strategies	30.6% (10.5)	4.5% (1.9)	0.2% (0.2)	0.0% (0.0)	1.2% (1.0)	0.0% (0.0)	4.6% (2.5)	0.7% (1.3)	26.1% (9.8)	0.5% (1.1)	0.7% (0.6)	0.1% (0.2)	*63.3% (15.5)*	*5.7% (3.5)*
Judgment Strategies	7.5% (6.1)	43.2% (8.7)	1.8% (1.2)	4.1% (2.7)	3.9% (2.0)	7.7% (3.0)	9.1% (3.9)	19.6% (6.2)	13.5% (5.9)	15.7% (3.9)	0.9% (0.6)	4.1% (2.3)	*36.7% (15.5)*	*94.3% (3.5)*
Total	*38.1% (7.9)*	*47.6% (8.7)*	*1.9% (1.3)*	*4.1% (2.7)*	*5.1% (2.7)*	*7.7% (3.0)*	*13.6% (4.6)*	*20.3% (6.4)*	*39.6% (8.0)*	*16.2% (3.7)*	*1.6% (1.0)*	*4.1% (2.3)*	*100%*	*100%*

As can be seen from Table 6.2, which aggregates the decision-making behaviors in the two rating sessions, when rating the student scripts independently, the raters spent more energy interpreting the student scripts than judging them (M=63.3% interpretation strategies versus M=36.7% judgment strategies), whereas they spent far more energy judging the scripts than interpreting them in the communal rating sessions (5.7% interpretation strategies versus 94% judgment strategies). Table 6.1 shows that when interpreting the student scripts in the independent rating sessions, the raters mostly read or reread the student scripts and supported their image of the scripts by treating language errors and interpreting or discerning the ideas in the scripts, although contextual factors like envisioning the personal situation of the students and the task requirements were also taken into consideration. Having already created an image of the scripts in the independent rating sessions, the raters spent little time on interpreting them in the communal ratings, although at times they felt the need to look back and double check this image, as the following excerpt illustrates (reproduced from Chapter 5):

> *Jesper*: godt, så er det nummer 2. og det er den der hedder WORK ETHICS til overskrift.
> *Pernille*: der er halvanden side, ja.
> *Jesper*: ja
> *Pernille*: altså, den har jeg givet et 8-tal for.
> *Jesper*: så er vi på den, Pernille.
> *Pernille*: nå
> *Jesper*: jeg har 5 til 6.
> *Pernille*: er det rigtigt?
> *Jesper*: ja, ha, ha
> *Pernille*: ha, ha. åh Gud. Så er jeg næsten nødt til at læse den igennem. Skal jeg ikke?
> (Pernille and Jesper, script 2)
>
> *Translation:*
> *Jesper: well, number 2. And it is the one with WORK ETHICS as the title.*
> *Pernille: it is a page and a half, yes.*
> *Jesper: yes*
> *Pernille: actually I have given it an 8.*
> *Jesper: so we've got a situation, Pernille.*
> *Pernille: oh.*
> *Jesper: I've got 5 to 6.*
> *Pernille: is that right?*
> *Jesper: yes, ha, ha.*
> *Pernille: ha, ha, oh my God. Then I'll have to read it again. Shouldn't I?*
> *(Pernille and Jesper, script 2)*

When judging the student scripts in the independent rating sessions, the raters attended to a variety of textual factors (mainly language related features) and monitored themselves mainly by articulating scores, articulating general impressions and even noting concrete examples directly from the scripts, as a strategy for preparing themselves for the communal rating sessions to come (excerpt reproduced from Chapter 5):

> Gælder om at tage nogle notater til den fælles evaluering.
> (Tina, script 1)
>
> *Translation:*
> *It is all about taking notes in preparation for the communal rating session.*
> *(Tina, script 1)*

The scores articulated in the independent rating sessions were often flexible scores (e.g. "a high 8" or a "5 maybe 6"), indicating that the scores given in the independent rating sessions were open for negotiation (see Appendix H for full range of scores).

The monitoring and textual foci while judging the student scripts were upheld in the communal rating sessions, although the raters here reduced their attention to language related features to make room for other textual features of the scripts. This textual focus was in the communal ratings supported by monitoring themselves, their co-raters or consulting the context. As in the independent rating sessions, the raters here articulated their scores, defined, revised or suggested assessment strategies and exemplifyed directly from the student scripts, implying a joint effort between the raters to justify their scores and improving their assessment strategies. As in the independent rating sessions, the raters would not directly consult the scale descriptors but chose to rely on each other for resolution.

6.3 Sequence of Decision-Making Behaviors

The sequence of the raters' decision-making process was traced from their independent rating sessions to their communal rating sessions. It was shown in Figure 4.1 (Chapter 4) that the raters went through three phases in the independent rating sessions where they scanned the scripts to

form a preliminary image, thoroughly interpreted and judged the scripts, and finalized a preliminary score in preparation for the communal rating sessions. Figure 5.1 (Chapter 5) showed that when meeting with their co-raters in the communal rating sessions, the raters also went through three phases to come to a final conclusion on a score. They exchanged preliminary scores from their independent rating sessions, exchanged and refined their judgments, and finally assigned a score. The rating sequence from the independent rating sessions to the communal rating sessions is illustrated in Figure 6.1.

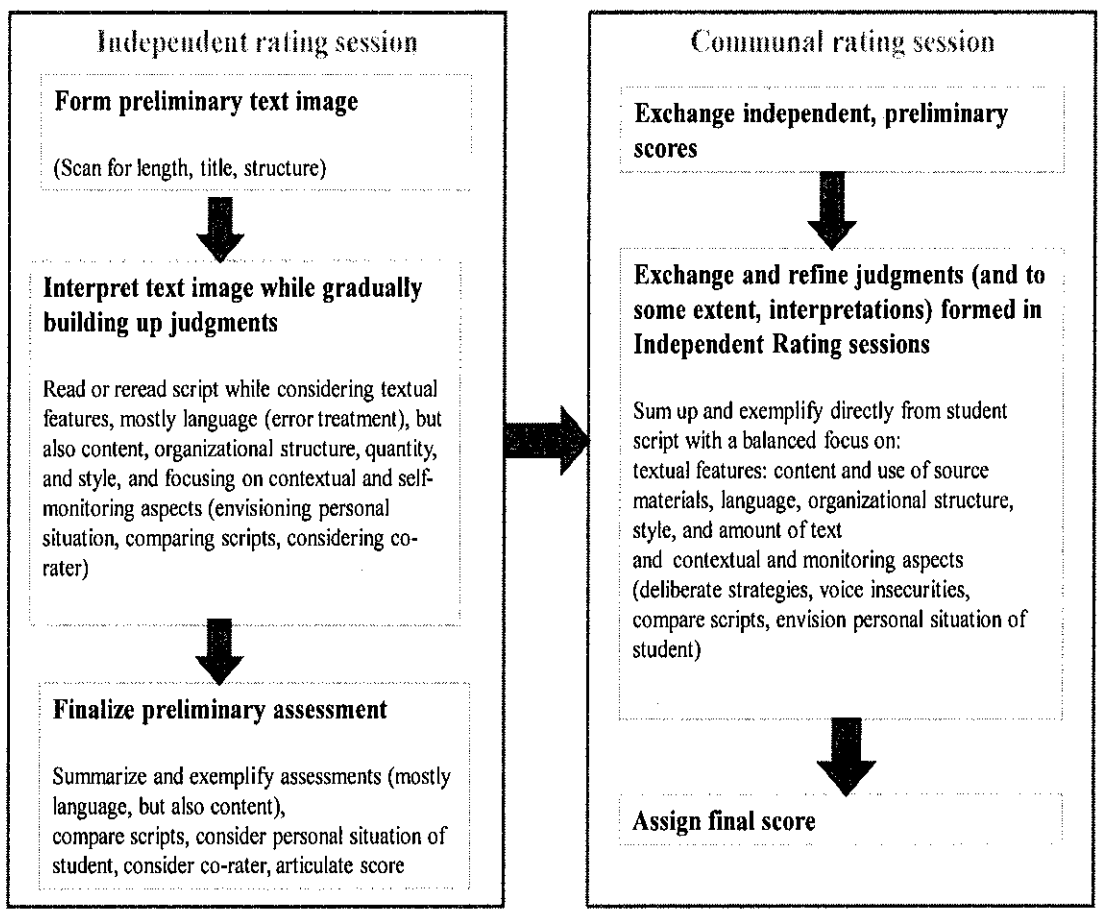


Figure 6.1: Type and Sequence of Raters' Decision-Making Behaviors from Independent Ratings to Communal Ratings

6.4 Balance of Attention to Official Assessment Criteria

The progression of the raters' distribution of attention to textual features corresponding to the official assessment criteria in the independent ratings and in the communal ratings is illustrated in Figure 6.2 (reproduced from Figures 4.2 and 5.2).

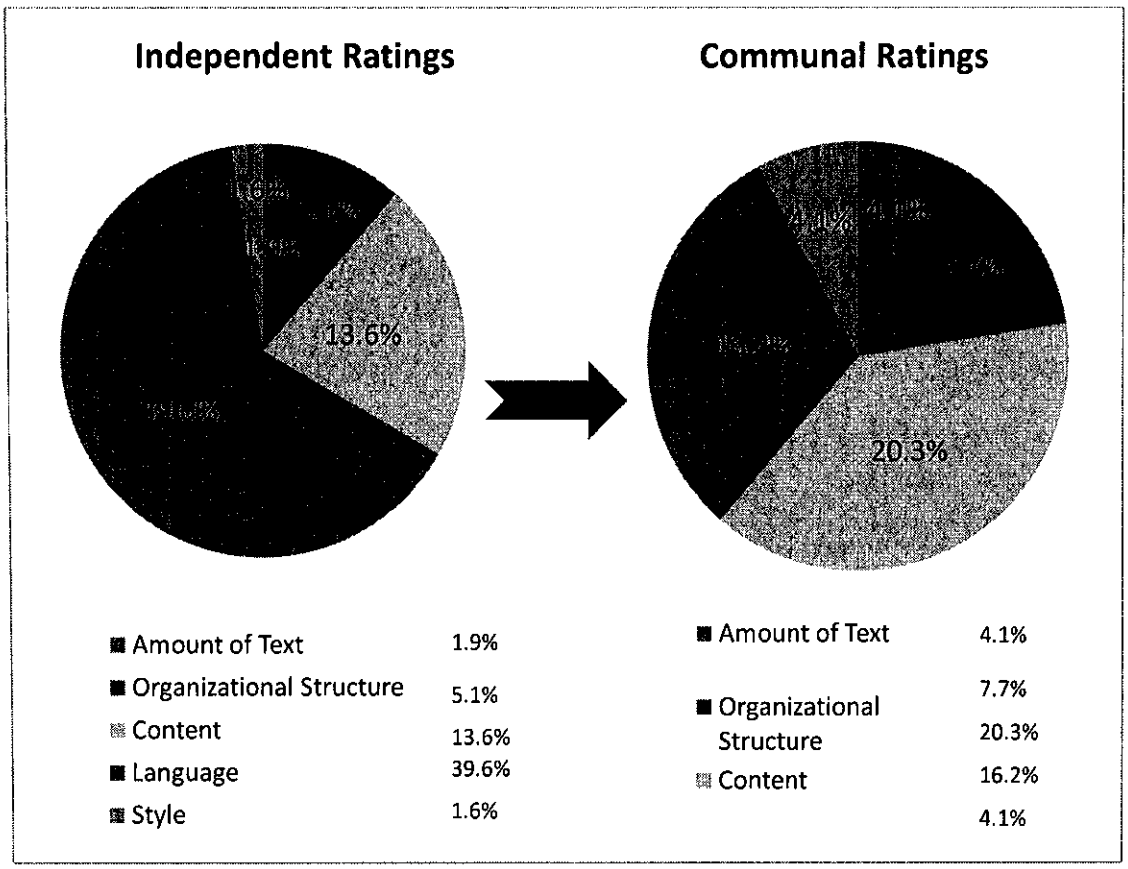


Figure 6.2: Mean Percentages of Textual Features Attended to from Independent Ratings to Communal Ratings

As can be seen from Figure 6.2, when moving from the independent ratings to the communal ratings, the distribution of attention to the textual features corresponding to the official assessment criteria became more balanced: in the independent ratings the raters paid an excessive amount of attention to language. In the communal ratings, however, this attendance to

Language was reduced to more than half, leaving more room to attend to Content in particular, but also to other aspects of the student script: Organizational Structure, Style and Format, and Amount of Text. This tendency in the communal ratings to reduce the attention to language-related features of the scripts to make room for a more balanced attention to the different textual features was emphasized in the reflective reports of the raters' perceptions of CWA (discussed below and seen in Appendix I):

> Alle aspekter (idiomatik, grammatik, stilistik, indhold) får mulighed for at blive inddraget (Henrik, retrospective report)
>
> *Translation:*
> *All aspects (idiomatic, grammar, content) will be assessed.*
> *(Henrik, retrospective report)*

A distinction was made in the coding scheme between interpretation strategies and judgment strategies, which facilitated an analysis of how the raters distributed their attention to the different textual features when interpreting the scripts and when judging them. Results showed that the distribution of attention to the textual features became more balanced (with far less attention to language-related features) as the raters judged the script images compared to when they interpreted the scripts. When placing the interpretation strategies and the judgment strategies alongside the communal rating strategies (Figure 6.3), we see that as the raters moved from interpretation strategies to judgment strategies in the independent rating sessions and further to the communal rating sessions, the distribution of attention to textual features corresponding to the official assessment criteria became progressively more balanced as less attention was paid to language and more room was left for attention to other textual features.

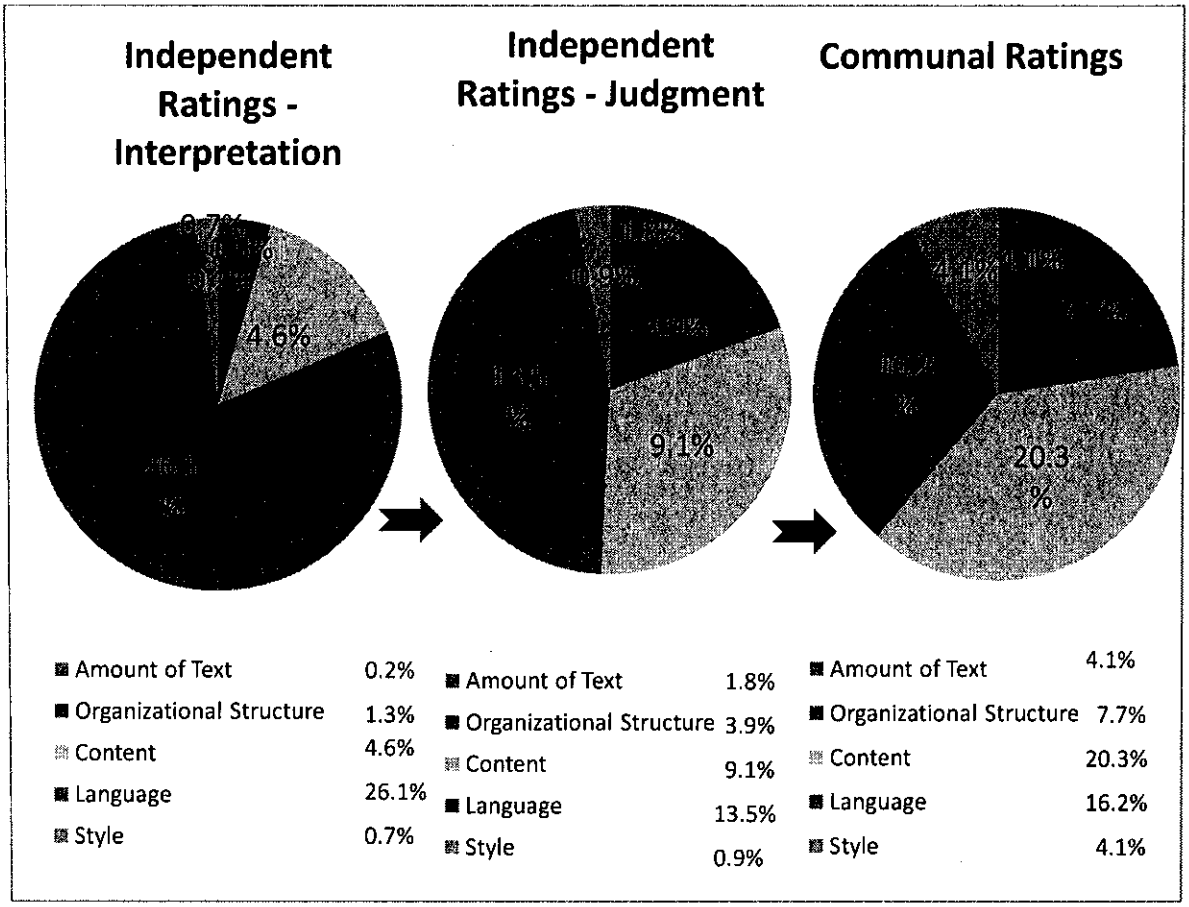


Figure 6.3: Mean Percentages of Textual Features Attended from Interpretation Strategies to Judgment Strategies in Independent Ratings to Communal Ratings

6.5 Distribution of Scores

The findings on the raters' agreement levels in the independent rating sessions and in the communal rating sessions are reproduced here, supported by a presentation of the raters' score range in the two sessions.

6.5.1 Rater Agreement

The independent ratings showed a rater agreement level of .87 (Kendall's *W*) and .75 (Cronbach's Alpha), and the communal ratings displayed an agreement level of .90 (Kendall's *W*) and .80 (Cronbach's Alpha). There was little if any difference between the levels of agreement among raters in either method of (independent or communal) scoring, as evidenced by Kendall's *W* or by Cronbach's alpha. The lack of increase in rater agreement from the independent ratings to the communal ratings is further reflected by the little difference in means and standard deviations in the two rating sessions. Table 6.3 shows, for the 15 student scripts, the means and standard deviations for the 20 raters' scores in the independent ratings and the 10 dyad ratings of the 150 scores in the communal ratings.

Table 6.3: Means (M) and Standard Deviations (SD) of Independent Scores and Communal Scores (20 raters rating 15 scripts=300 scores in independent ratings and 10 rater dyads rating 15 scripts=150 scores)

Scripts	Independent Ratings	Communal Ratings
1	6.3 (0.6)	6.5 (0.7)
2	7.9 (0.7)	8.0 (0.7)
3	7.9 (0.9)	7.9 (0.9)
4	9.1 (0.6)	9.2 (0.4)
5	8.0 (0.9)	7.7 (0.5)
6	9.0 (0.7)	9.0 (0.7)
7	9.1 (0.6)	8.9 (0.6)
8	6.5 (1.0)	6.3 (1.1)
9	5.0 (0.5)	4.8 (0.4)
10	4.7 (0.7)	4.5 (0.5)
11	4.7 (0.5)	4.6 (0.7)
12	8.1 (1.0)	8.3 (1.1)
13	8.7 (0.6)	8.7 (0.7)
14	5.1 (0.6)	4.9 (0.6)
15	6.1 (0.8)	5.8 (0.8)
Average	7.1 (0.7)	7.0 (0.7)

6.5.2 Score Range

A closer look at the distribution of scores in the two rating sessions (Table 6.4) reveals that behind these seemingly similar rating sessions, there was a difference in the range of scores assigned to the scripts.

Table 6.4: Range of Scores in Independent Ratings and in Communal Ratings (20 raters rating 15 scripts=300 scores in independent ratings and 10 rater dyads rating 15 scripts=150 scores)

	Independent Ratings				**Communal Ratings**		
	Minimum score	**Maximum score**	**Score range**		**Minimum score**	**Maximum score**	**Score Range**
1	5.3	7.5	2.2		5	7	2
2	5.5	8.7	3.2		7	9	2
3	6.7	9.7	3		7	9	2
4	8.0	10	2		9	10	1
5	6	10	4		7	8	1
6	8	10	2		8	10	2
7	8.3	10.3	2		8	10	2
8	5	8	3		5	8	3
9	4	6	2		4	5	1
10	4	6.5	2.5		4	5	1
11	4	6.7	2.7		4	6	2
12	6.0	10	4		7	10	3
13	8	10	2		8	10	2
14	4	6	2		4	6	2
15	5	7.5	2.5		5	7	2
Average	5.9	8.5	2.6		6.1	8.0	1.9

As can be seen from Table 6.4, there was a reduction in the range of scores for most individual student scripts when going from the independent ratings to the communal ratings. With a 10 point scale, the independent ratings had an average score range of 2.6 compared to 1.9 in the communal ratings. The raters differed by 2 or more points in the independent ratings, sometimes varying by 3 or 4 points. In the communal ratings, the raters disagreed by 3 points on only two of the scripts, and differed by 1 or 2 on the rest of the scripts. The difference in the score range was most noticeable in the judgment of script 5: receiving scores from 6 to 10 in the independent ratings and scores from 7 to 8 in the communal ratings. This script reduced its score range from 4 to 1 when going from the independent ratings to the communal ratings. The independent ratings in this study may have displayed a wider score range than the communal ratings as there

were twice as many independent scores (n=300) as communal scores (n=150). What we can conclude, however, is that communal ratings reduced that range of scores slightly, particularly by getting rid of the most extreme scores. This benefit was reflected in the retrospective questionnaire in which a number of raters commented that CWA managed to weed out idiosyncratic and mistaken judgments (see below and Appendix I for rater comments).

6.6 Raters' Perceptions of CWA

To elicit their opinions of the whole CWA procedure from the independent ratings to the communal ratings, the experienced raters in this study were asked retrospectively about their perceptions of score accuracy in the two rating sessions and about their perceptions of CWA practices in general. Although their comments might be colored by their loyalty to their profession as CWA raters of the national rater corps, these raters were highly experienced, so their perceptions of CWA can provide expert insight into the merits of such assessment procedures.

6.6.1 Perceptions of Score Accuracy

In a retrospective questionnaire the raters were asked to reflect on the cases of score discrepancy and to state which score (their original, independent score or the final, communally rated score) they found more accurate. Table 6.5 shows the 20 raters' perceptions of accuracy of scores in the two rating conditions.

Table 6.5: Raters' Perceptions of Score Accuracy

Number of discrepant scores[47]	160
Number score discrepancy cases in which raters believed their independent score to be more accurate	8 (5%)
Number of discrepancy cases in which raters believed the final, communally rated score to be the more accurate	152 (95%)

In almost all (95%) cases of score discrepancy the raters believed the communally rated score to be more accurate than their own independent score. This indicates a strong faith in CWA and implies that the raters were willing in this context to modify their assessments of the particular student scripts, and perhaps their general assessment strategies as well.

It must be noted, however, that what could contribute to this almost unanimous agreement that the communal rating sessions produced more accurate scores than the independent rating sessions was the fact that the raters participating in this study were all part of the national rater corp. By expressing confidence in CWA they were in a sense validating their jobs as CWA raters. Also, it is likely that the raters felt more committed to the communally rated scores because they were their most recent decisions.

6.6.2 Raters' General Perceptions of CWA

In the retrospective questionnaire the raters were also asked what they perceived to be the advantages or disadvantages of co-rating procedures in general. All 20 raters answered the questionnaire. Their comments were all positive and fell into the following broad categories as shown in Table 6.6 (a full list of the raters' responses can be seen in Appendix I):

[47] Adjacent scores (such as '8' and '9') were also considered discrepant.

Table 6.6: Raters' General Perceptions of CWA

Perceptions of CWA (20 raters)	**Number of raters who gave such a comment**
They offer the best opportunity to reach the most accurate score possible	17
They offer the raters an opportunity refine their assessment strategies	4
They ensure that raters assess by the same standards	5

Most of the raters (17) regarded the main advantage of CWA procedures to be the opportunity to reach the best possible scores. Of these many commented that "fire øjne er bedre end to" (= "four eyes are better than two") (Gitte, Tina, Jens, Jette, Ken). More specifically, some raters mentioned that the reason that CWA opens up an opportunity to reach the most accurate score possible is that the discussions weed out idiosyncratic assessments:

> At ens egne idiosynkrasier/foretrukne ting ikke får for meget vægt
> (Jesper)
>
> *Translation:*
> *That one's own idiosyncrasies/preferred aspects don't carry too much weight*
> *(Jesper)*

And they prevent mistaken judgments from counting towards final scores:

> det sker da at man med en enkelt opgave får revideret sin bedømmelse, for der er ting man totalt har overset, måske er blevet forstyrret
> (Susanne)
>
> *Translation:*
> *it does happen that one has to revise one's assessment because one has overlooked things, was distracted*
> *(Susanne)*

One rater also believed the communal ratings made sure that too much focus was not put on language-related features:

> En anden ting er vægtningen af sprog og indhold i forhold til hinanden. Man kan fokusere så meget på gram.fejl, at man faktisk glemmer, at eleven har forsøgt at formulere nogle komplicerede sætninger indholdsmæssigt, så der kommer ekstra mange fejl.
> (Gitte)
>
> *Translation:*
> *Another thing is the importance attached to the content and language of the scripts. You can focus so much on grammatical errors that you actually forget that the student has attempted to construct complex sentences about some idea, so there will be a lot of extra errors.*
> *(Gitte)*

Four raters commented on the opportunities CWA offers for refining and reassessing general assessment strategies, indicating that there is a rater development potential inherent in such assessment practices. As one rater put it:

> vi [har] alle brug for at prøve vores bedømmelser af med andre
> (Susanne)
>
> *Translation:*
> *we all need to test our judgments against others*
> *(Susanne)*

Five raters mentioned that CWA makes sure that raters assess by the same standards. One could argue that conscious matching of the scripts to the rating scale would result in assessments by the same standards. However, as one rater pointed out, such a matching exercise is close to impossible.

> Det er vigtigt, fordi censorer skel helst have så meget fælles grundlag som muligt, for der er mange ting man ikke kan opstille regler for. Man kan ikke sådan bare give et tal. Derfor er samtale vigtigt
> (Pernille)
>
> *Translation:*
> *It is important because raters must judge by the same standards. There are so many things you cannot make rules for. You cannot just assign a score like that. That is why conversation is important*
> *(Pernille)*

As with the perceptions of the accuracy of the scores assigned, the raters' comments on the CWA practices in general may have been overwhelmingly positive because the raters are

members of the national rater corps. Their positive attitude to CWA could be a way of validating their jobs as national raters.

Despite the probability that the raters' perceptions of CWA might be colored by their loyalty to their profession as CWA raters, CWA was perceived to hold strong advantages. Not only was CWA perceived to produce more accurate scores because idiosyncratic prejudices can be illuminated, they were also perceived to hold a rater development potential in that raters' here get a chance to validate their assessments with and against other professionals. Further, such procedures were believed to secure a national framework for writing assessment.

6.7 Summary of Raters' Decision-Making Behaviors from Independent Ratings to Communal Ratings

As they moved from the independent rating sessions to the communal rating sessions, the raters in this study gradually refined their assessments to reach a final score for each script. In the independent rating sessions they carefully interpreted the student scripts while gradually building up judgments to reach a preliminary score in preparation for the communal rating sessions. In the communal rating sessions the raters exchanged and validated their assessments to reach what they believed were the most accurate scores.

The raters displayed the same types of strategies in the two rating sessions. However, the distribution of these behaviors varied to some extent as they progressed from the independent rating sessions to the communal rating sessions. From interpreting to forming their judgments of the student scripts in the independent ratings, to refining and finalizing these assessments in the communal rating sessions, the raters progressively reduced their attention to language-related features of the scripts, leaving more room for attention to other textual features such as content and use of source materials in particular, but also to organizational structure, style/format and amount of text, thereby balancing their attention more evenly to the textual features corresponding to the official assessment criteria.

There was little, if any, difference between the levels of agreement in the independent ratings and in the communal ratings. Many extreme scores appearing in the independent ratings were, however, eliminated during the communal ratings.

Although caution should be taken about endorsements of CWA by raters from the national rater corps, the retrospective reports evidenced an overwhelmingly positive view of CWA. The raters almost unanimously agreed that the communal ratings produced more accurate scores, and in general they perceived this rating procedure to be maintaining national standards as well as providing them with an opportunity to develop professionally as raters.

Chapter 7
Discussion and Conclusion

7.1 Purpose and Scope of the Chapter

This chapter concludes the study of raters' decision-making behaviors in CWA. It discusses the results and relates them to previous relevant research. Chapter 4 mapped the raters' decision-making behaviors when they rated student scripts independently. Chapter 5 mapped the raters' behaviors when they collaborated with another rater to reach a final score for each student script in the communal rating sessions. Chapter 6 traced the progression of the raters' decision-making behaviors by comparing their behaviors in the independent rating sessions with their behaviors in the communal rating sessions. The results from these chapters are discussed here in relation to the research questions posed in chapter 2 and to previous research into the field of raters' decision-making behaviors and communal rating practices. The chapter further discusses the validity implications of CWA and implications regarding the concept of rater expertise. Finally, the chapter concludes by relating the present study to the Danish context and by suggesting further research.

7.2 Summary of Findings

In the CWA practice studied here, in which raters first assign preliminary scores in independent rating sessions and afterwards meet face-to-face with another rater in communal rating sessions to finalize a score for each script, the raters went through a rating process of conscientiously creating an image of the student scripts while gradually making their scoring judgments. As they progressed from one rating session to the other, the raters diligently validated their perceptions of the scripts and balanced their attention to the official assessment criteria more evenly and arrived at what they believed to be the most accurate scores for each script. If problems or insecurities arose the raters chose not to consult the scoring rubrics directly, but preferred to rely on the professional interactions with other raters. What exactly went on in the two rating sessions and

how the raters progressed on their decision-making behaviors from one session to the other is summarized and discussed below while addressing each research question stated for this study:

7.2.1 What are the raters' distinct decision-making behaviors and what is the sequence of these behaviors in the independent rating sessions and in the communal rating sessions?

The raters in the present study typically went through a sequence of three phases in both rating sessions. In the independent rating sessions they would first go through a phase of creating a preliminary image of the student scripts by scanning for length, title and, to some extent, organizational structure in order to create a preliminary image of the scripts. Next they would enter a phase of carefully constructing an image of the scripts by reading them, interrupting their reading process by treating errors and discerning or interpreting the content of the scripts and gradually building up their judgment of the scripts. In the last phase of the independent rating sessions the raters would sum up their assessments to finalize a preliminary score in preparation for the communal rating sessions to come. When unsure of which score to assign they would never directly consult the scoring rubric but would revisit their impressions, compare student scripts and consider or conjecture on their co-rater's assessments. The raters were very conscious of the subsequent communal rating sessions as a forum available to validate and finalize their scores in that they would often assign flexible scores and explicitly mention that they needed help from their co-rater to validate and finalize their assessments, and they would make specific and concrete examples in preparation for the negotiations in the rater discussions. This type and sequencing of behaviors mirrors the complex behaviors identified by other studies into the rating process (e.g. Cumming et al. 2001, 2002; Lumley, 2005; Milanovic et al. 1996; Sakyi, 2003). However, unlike the rating sessions in these studies, the independent rating sessions in the present study were followed by communal rating sessions, and this apparently had an effect on the raters' decision-making in that the raters were very much aware of the opportunity given to them to have their scores resolved in the succeeding communal rating sessions. Although other studies (e.g. DeRemer, 1998; Lumley, 2002; 2005: Sakyi, 2000, 2003; Vaughan, 1991) have found that raters do not match scripts directly to the scoring rubrics, raters in these other studies

did at times consult the rubrics either for score resolution or as a way of articulating their judgments (Lumley, 2002, 2005). The raters in the present study never directly consulted the scoring rubric although it was available to them, seemingly because they believed interactions with other professionals later would help them assign more accurate scores.

The communal rating sessions also consisted of three phases although the second phase would typically be omitted if the raters had assigned identical scores in their independent ratings. First the two raters would exchange their scores from the independent rating sessions, sometimes combined with a brief explanation for their scores. If the raters' independent scores were the same, the raters would head straight to the final phase. If there was a discrepancy in the raters' independent scores, they would go through a second phase, where they would deliberate and validate their scores and assessment strategies against each other by making explicit their specific assessment strategies, by exemplifying directly from the student scripts and by revisiting their notes and sometimes the student scripts during the process. They would here also articulate their general impressions, compare scripts and envision the personal situation of the students. This validation process would lead into the third and final phase, where the raters finalized their scores. If after careful deliberation, the raters were still unsure of which exact score to assign, they never directly consulted the scale descriptors, but rather reconsidered their assessments. The raters' process of carefully deliberating and refining their assessment strategies confirms previous claims of CWA's potential for rater development seen in other studies on communal ratings (e.g. Allen 1995; Durst et al. 1994; Moss et al., 1998; Nixon & McClay, 2007). The results further echoes the sequence of behaviors identified by Moss et al. (1998), who investigated communal rater behavior for math teacher licensure. However, unlike the raters in their study, the raters in this study revisited the student scripts and made many direct references to the scripts for the purpose of presenting concrete evidence for their assessment claims. The reason for this difference might be that unlike the raters in Moss et al.'s (1998) study, the raters in this study were experienced in communal rating practices and thus might have learned that exemplifying directly from the student scripts is a useful negotiating technique.

7.2.2. How do the raters distribute their attention to the official assessment criteria in the two rating sessions and how does this distribution of attention differ from one session to the other?

The official assessment criteria for the written EFL exam investigated in this study were Amount of Text, Organizational Structure, Content and Use of Source Materials, Language and Style/Format. The raters' distribution of attention to the textual features corresponding to these assessment criteria varied from the independent rating sessions to the communal rating sessions and from the raters' interpretation strategies to their judgment strategies within the independent rating sessions.

During the process of creating an image of the student scripts (interpretation strategies) in the independent rating sessions, the raters paid an excessive amount of attention to Language, but would reduce this attention when they judged this image. This attention to Language was reduced even further in the communal rating sessions, leaving the raters an opportunity to balance their attention more evenly among the textual features corresponding to the official assessment criteria, especially textual features related to Content and Use of Source Materials. So, even though the raters are not instructed to pay an equal amount of attention to the textual features corresponding to the official assessment criteria of Amount of Text, Structure, Content and Use of Source Materials, Language, and Style/Format, the communal rating sessions certainly prompted the raters to distribute their attention more evenly across those criteria. The finding that attention to language does not overshadow attention to other textual features in CWA does not echo Broad's (2003) finding that CWA raters focus more on mechanics than on anything else because such features are "safer to talk about" (Broad, 2002:63). This difference in the level of attention to language or mechanics in the two studies might be explained by the raters' level of experience: unlike the raters in Broad's (2003) study, the raters in this present study were experienced CWA raters and might have learned by experience to balance and articulate their attention to a broader spectrum of textual features.

The development towards a more evenly distributed attention to textual features corresponding to the official criteria (or, one could say, a decrease in language focus) in this study could mean:

1. The raters, reading unfamiliar students' texts for the first time containing many errors in English, have to spend considerable time struggling to interpret these language errors not just to make sense of the texts, but also to comprehend the nature and score of the errors in English in order to reach a summary judgment of the language abilities demonstrated.

2. When judging the images they have created of the student scripts, the raters in the independent rating sessions monitor themselves to prepare for the communal rating sessions. Perhaps it is perceived as more "politically correct" among the raters in this rating community to place their attention relatively evenly across the official criteria (or to pay less attention to language), and being aware of this tendency in the rating community, the raters manage to weed out their language-focused attention in their independent ratings, thereby monitoring themselves towards the rating community standards expressed in the communal rating sessions.

3. There is a clear distinction between interpretation strategies and judgment strategies (confirming studies by Cumming, 1990; Cumming et al.'s 2001, 2002 and Erdosy, 2004). The raters use interpretation strategies to create an image of the student scripts, and they use judgment strategies to judge that image. Thus, when trying to figure out how raters weigh or assess the different textual features of a script (and/or the corresponding official criteria), it is imperative to make a distinction between interpretation strategies and judgment strategies. This poses a challenge to studies in which rater comments are counted in order to find out what textual features raters weigh in their assessments of student scripts (as was done, for instance, in Milanovic et al, 1996; Vaughan, 1991; Wolfe et al., 1998), because although counting rater comments may help us understand what features raters attend to for assessment purposes, these comments may not reveal how raters choose to weigh these features when judging the scripts.

7.2.3 To what extent do the raters engage equally in the communal rating sessions?

Moss (1996:26) argued that, in order for communal assessment to be successful, raters must remain equally engaged in the rater discussions. In other words, one rater must not always be the dominant rater and the other rater the conceding one. Although Johnson et al. (2005) interpreted this in terms of score dominance, Moss herself and colleagues (Moss et al., 1998) viewed it from the perspective of conversational dominance. The present study investigated score dominance and conversational dominance as well as the interaction between these two types of dominance, and the retrospective questionnaire served to shed further light on this issue.

The raters in this study exhibited some degree of score dominance, although not to a large extent. In cases of a discrepancy between the two independent scores the final score would sometimes be a compromise, i.e. the final score would fall midway between the two scores. At other times, however, it would be a case of score concession, i.e. the final score would be closer to the independent score of one rater than the independent score of the other rater. Thus one rater in the rater dyad would concede his/her score while the other one would dominate by score. Although no rater exhibited score dominance in all score concession cases, three raters out of 20 demonstrated a tendency to dominate by score. This score dominating tendency was not reflected in any notable signs of conversational dominance in that there was no major difference in the number of words or decision-making behaviors that each rater in the rater dyads contributed to the conversations. In fact, the one rater that produced more than twice as many words and decision-making behaviors than his co-rater, tended to concede his scores. An examination of the relationship between score dominance and conversational dominance in all concession cases showed that score dominance was not related to conversational dominance. The raters would produce on average 216.9 words if they conceded their scores and 204.6 words of they dominated by scores. Similarly, they would produce on average 11.2 decision-making behaviors when conceding and 9.9 decision-making behaviors when dominating by score. This implies that score dominance was not a product of conversational dominance and indicates that the communal rating sessions were characterized by equal engagement and not by an atmosphere of suppression. The retrospective questionnaires support this claim in that when asked in cases of

score discrepancy which score they found more accurate, their own independent score or the final, communally rated score, the raters preferred the final, communally rated score in almost all (95%) cases.

In short, these findings showed that communal ratings produced an opportunity for the raters to engage equally in a rating discussion to reach what each believed to be the most accurate score for each of the student scripts. This confirms Moss et al.'s (1998) and Johnson et al.'s (2005) findings that dominance is not a major issue in communal ratings. The findings in the present study further demonstrated that although some raters tend to dominate by score, score dominance does not seem to be a product of conversational dominance. This refutes Johnson et al.'s (2005) assumption that score dominance indicates that raters are not equally engaged in rater discussions and thus that score dominance invalidates the final scores. If after validating their scores against other raters, raters agree that the final scores are more accurate than their own independent scores, it seems that score dominance may not invalidate the final scores.

7.2.4 What are the agreement levels and score ranges in the two rating sessions and how do they compare to one another?

The score agreement levels and the score ranges were calculated in both rating sessions. The independent rating sessions showed an agreement level of .87 (Kendall's W) and .75 (Cronbach's Alpha), and the communal rating sessions displayed an agreement level of .90 (Kendall's W) and .80 (Cronbach's Alpha). Thus, there was little, if any increase in the score agreement levels in the two rating sessions. Being a high stakes test these agreement levels might be considered to be at the low end of the acceptability level: Johnson et al. (2005) contend that "research studies and low-stakes assessments require a minimal reliability of .70, whereas in applied settings with high stakes tests generally require a minimal reliability of .90" (2005:132). The Norwegian[48] Ministry of Education suggests that a minimum reliability level of .80 (Cronbach's Alpha) be required (and a minimum of .85 desirable) in evaluating students'

[48] The Danish Ministry of Education has not issued such guidelines, which is why I refer to a country whose educational system is comparable to the one in Denmark. The source is taken from www.utdanningsdirektoratet.no (Rammeverk for nasjonale prøver 2007).

writing. Coffman (1971), however, warned against setting equal reliability standards across subjects, indicating that reliability naturally decreases as the subject area becomes fuzzier and less definable, with mathematics having the highest levels of reliability and subjects like history and composition holding lower levels of reliability. Supporting this argument, Penny, Johnson, and Gordon argued that "The level of specificity attained in the sciences and in mathematics ... is more difficult to achieve in subject areas in which diversity in responses is accepted, and even valued" (2000:271), suggesting that composition, by its less definable nature faces a smaller chance of high reliability scores. Breland (1983) even found a difference in reliability levels depending on scoring procedures, reporting higher inter-rater reliability levels in analytic scoring than in holistic scoring. With the HHX being an integrated, performance-based writing test using holistic scoring, one could say that the relatively low agreement level would be acceptable.

The score range decreased as the raters moved from the independent rating sessions to the communal rating sessions. Not only was the average score range reduced from 2.6 to 1.9, but extreme score ranges were reduced dramatically as evidenced by one script, which received scores from 6 to 10 in the independent rating sessions and only the scores 7 or 8 in the communal rating sessions. Although a decrease in the score range is to be expected because there were only half as many communally rated scores as independently rated scores, it was clear that the communal rating sessions managed to get rid of the most extreme scores. This advantage of the communal ratings was reflected in the retrospective questionnaire, in which a number of raters commented that CWA often weed out idiosyncratic and mistaken judgments.

7.2.5 What are the raters' perceptions of CWA in general and in relation to the specific CWA they have just practiced?

The raters were asked retrospectively to report on their perceptions of CWA both in general and in relation to the specific CWA they had just practiced. With respect to their perceptions of the CWA they had just practiced the raters agreed almost unanimously that the communal ratings produced more accurate scores. Their strong belief in the validity of the CWA was further evidenced in their perceptions of CWA in general. They all viewed CWA in a positive light

arguing that CWA produced more accurate scores by eliminating mistaken or idiosyncratic assessments, that they helped maintain national standards, and that they provided the raters with an opportunity to develop professionally. Although the raters' positive views of CWA were probably colored by the fact that the raters were part of the national rater corps, their comments were overwhelmingly positive. This reflects rater views found in Moss et al.'s study (1998) in which raters believed that communal ratings empowered them to become better raters and better teachers. The findings, however, do not reflect the frustration the raters in studies by Broad's (2000) and Mohan and Low (1995) experienced when working with other raters in standard setting practices. The reason that the raters in the present study did not share the frustrations experienced by Broad's (2000) and Mohan and Low's (1995) raters might be that unlike those raters, the raters in this study were very experienced in CWA and thus might have accepted variation and disagreements as a natural characteristic of a rating community and of human judgment in general.

7.3 Validity Implications

In a test validation study one addresses the question of how valid the assessment practice under investigation might be. Validity is a complex matter, and the question of how valid such a practice is depends on what perspective is taken. A measure is often regarded as valid "if it does what it is intended to do" (Davies et al. 2002:221), and if we can confidently draw inferences about students' (writing) abilities on the basis of their test scores, with considerations of the consequences of the test (Bachman, 1997; Messick, 1989). A strong psychometric view of validity focuses on the extent to which assessments correspond to an absolute standard (e.g. Wolfe et al. 1998). Such a standard may be determined by the level of agreement among raters and/or the extent to which raters judge the scripts according to a set standard, often reflected in a prescribed rubric. From a postmodern, hermeneutic perspective, however, absolute standards do not exist, and validity does not necessarily depend on score agreement and absolute standards, but rather on the extent to which raters engage equally and critically (e.g. Moss, 1994, 1996) with other raters in a validation process where multiple perspectives are scrutinized, challenged and refined to reach what the raters believe is the most accurate score.

7.3.1 A Psychometric View of CWA

A psychometric view of writing assessment values consistency and conformity to set standards, and thus a psychometric approach to validating CWA would look at the extent to which raters in this study agree with each other and conform to the prescribed scoring rubric. From such a perspective CWA may not seem to be worth the additional resources and costs it requires over independent ratings.

7.3.1.1 Rater Agreement

Rater agreement levels have long been psychometricians' preferred means of validating a writing assessment practice. With little difference in levels of agreement among raters when moving from independent to communal ratings, psychometricians might find the communal ratings in the present study offered little increase in value. So communal ratings may not seem to be worth the resources they require. Nonetheless, the CWA procedures did, from a psychometric perspective, eliminate many of the extreme scores evident in the independent ratings and thus promoted conformity and consistency in this aspect of the scoring.

Lack of distinct differences in agreement levels does not necessarily indicate that communal ratings do not enhance rater agreement levels in general. The explanation for the lack of increase in agreement levels in the communal ratings of this study is very likely that the raters were highly experienced communal raters rating in this particular rating community: they had many years of experience rating with different raters in this community and thus were trained to rate similarly even in the independent rating sessions. In support of this, the findings from the present study showed that raters appear to monitor themselves toward the communal rating sessions, paying progressively more attention to other features besides language when judging the images they create of the student scripts, thereby preparing themselves to conform to the norms of their rater community. In this sense, the CWA procedures train (or rather, develop) raters into

becoming more reliable over time. The communal rating sessions are not unlike the training sessions set up to prepare raters for their future rating jobs in regular, "non-CWA" procedures.

7.3.1.2 Use of Scoring Rubric

Another objection strong adherents of the psychometrics tradition would have to CWA based on the findings from the current study is that the raters did not seem to directly consult the scale descriptors of the scoring rubric. Because the raters did not confer with the prescribed scale descriptors, we cannot be sure that they rated according to the standards set for the test.

This tendency among raters to not make explicit use of the scale descriptors is not only characteristic of CWA practices. Other, more traditional writing assessment practices, in which experienced raters rate scripts independently, seem to paint a similar picture. Studies into such practices have shown that scripts are not matched to the scoring rubrics, and if the rubrics are used at all, it is most often as an articulation tool through which raters justify their scores (see particularly, Lumley, 2002, 2005, but also DeRemer, 1998 and Sakyi, 2003). Matching a simplified, absolute scoring rubric to complex writer responses simply might not be possible. As one of the raters in the current study put it, "there are so many aspects you cannot make regulations for. You cannot just assign a score like that. That is why conversation is important" (Pernille). The present CWA procedures, however, appeared to differ from the traditional, non-CWA procedures by no direct reference at all to the descriptors. Knowing that the raters can have their assessments validated with another rater, they seemed to put more faith in the knowledge and expertise of the communal rating community than in the rubrics, including the scale.

It is, of course, difficult to tell whether the experienced raters in this study may have internalized the scale and therefore did not feel the need to consult it directly. However, when insecurities were expressed in the independent ratings, the raters did not consult the rubric, but rather expected their uncertainties about a score to be resolved in the succeeding communal rating sessions.

What did happen, however, in relation to the scoring rubric is that the textual features corresponding to the official assessment criteria were attended to, and when moving from the independent ratings to the communal ratings, the raters balanced their attention more evenly among these different features. Even though the guidelines make no recommendations as to how the raters should weigh the assessment criteria, the raters seemed to make sure that they were all attended to relatively evenly in the communal ratings.

7.3.2 A Hermeneutic View of CWA

A hermeneutic view of writing assessment would not attempt to validate CWA on the basis of rater agreement and conformity to a set standard, because an absolute score is not assumed to exist. The focus, rather, is on the raters' ability to make sound and contextualized assessments (Broad, 2000, 2003; Moss, 1994, 1996; Moss et al. 1998).

7.3.2.1 Sound Assessments

As Moss (1994, 1996) argues, a valid assessment is one which has been put to scrutiny by several perspectives, ideally one in which two or more raters engage equally and critically to refine their assessments and ultimately reach what they believe to be the most appropriate score. From this perspective, the CWA in the present study seemed to offer a fertile ground for cultivating valid assessments. Not only did the raters in this assessment practice get a chance to refine their judgments of the particular scripts in a current assessment situation, but they were also given the opportunity to refine their assessment strategies in general, thereby fostering sounder judgments in future assessment situations.

7.3.2.1.1 Sound Assessments of Current Scripts

The findings from the current study show that the raters in their CWA practice conscientiously went through a process of refining their judgments by engaging equally and critically in their assessments of student scripts.

7.3.2.1.1.1Equal Engagement

Their equal engagement was manifested in the extent to which the raters in this study contributed equally to the rater discussions. They contributed equally in terms of amount of words and decision-making behaviors. Further, when one rater dominated by score, this dominance was not a product of conversational dominance, but rather, it seems, of carefully deliberating different perspectives. The fact that in virtually all score discrepancy cases, the raters perceived the final score to be more accurate than their own independent score testifies to the claim that the raters' individual views were not suppressed.

7.3.2.1.1.2 Critical Engagement

Although it is difficult to operationalize the extent to which these raters engaged critically in their rater discussions, it can be argued that they put their assessments to scrutiny by defining/revising/suggesting their assessment strategies, by exemplifying directly from the student scripts, and in cases of a strong discrepancy between their independent scores by taking their time to revisit or even reread entire scripts.

7.3.2.1.1.3 Development of Assessments

What further testifies to a refinement of the raters' assessments in the present CWA practices was the development of rater focus from the independent ratings to the communal ratings. Moving from the independent ratings to the communal ratings the raters displayed a more balanced consideration of the textual features corresponding to the official assessment criteria, reducing their almost exclusive focus on language (during the independent ratings) to make room for other relevant textual features. Also, the raters' statements that they perceived the final,

communal scores to be more accurate than their own independent scores, mainly because mistaken and idiosyncratic assessments can be eliminated and precluded from counting towards the final score, testify further to the claim that the assessments had gone through a refinement process from the independent ratings to the communal ratings.

7.3.2.1.2 Sound Assessments of Future Scripts

This study cannot assert precisely how CWA may have directly affected the raters' general assessment strategies because the research did not trace their development longitudinally nor was it an experiment that compared conditions of communal ratings to a control group who did not perform CWA. But this thesis research does indicate some of the potential for rater development inherent in the CWA context. Raters in this study actively validated their assessment strategies against those of other, peer raters, deliberating over their assessment and voicing their mutual insecurities, and they ended up having a stronger faith in the communally rated scores than in their own independent scores. These trends point in the direction of CWA having the potential to refine progressively raters' assessment strategies and self-confidence. Further, when asked about their perceptions of CWA in general, a number of the raters explicitly pointed to the development potentials of such assessment practices, providing comments such as "we all need to test our assessments against others" (Susanne).

7.3.2.2 Contextualized Assessments

While psychometricians may be concerned that raters in CWA seemed to disregard the standardized scale descriptors in the scoring rubric, this may not as disheartening to those who take a more hermeneutic approach to writing assessment. In a hermeneutic view of assessment, no absolute textual value exists and thus the rubric that represents such a value may be close to futile or even damaging to the rating process because it forces the raters into rating unnaturally against their personal responses to students' writing. Reducing the role of the scale descriptors makes the assessments more contextualized in that it offers an opportunity to bring in different perspectives or responses to the particular scripts. The insecurities expressed by the experienced

raters in this study as well as their frequent inability to assign exact scores in the independent ratings give evidence about the inadequacy of scale descriptors and the need for contextualized assessments.

Moss (1996) suggests a dialectic between the standards and the local context, ideally a situation in which the standardized set (e.g. the scoring rubric) can be contextualized in particular assessment situations. In this regard the raters in this study did attend to relevant textual features corresponding to the official assessment criteria in the test. However, the scale descriptors in the rubric were largely ignored, and rather than taking a starting point in those descriptors and contextualizing them to the particular rating situation, the raters relied on the standards that had gradually been built up within the local rating community. Thus the assessments were contextualized and standardized with respect to the rating community rather than with respect to the scale descriptors.

7.3.3 Making Inferences and Observing Consequences of CWA

In terms of general validity concepts such as the extent to which the test measures what it is supposed to measure and whether we can infer a student's writing ability on the basis of his/her test scores as well as whether the consequences of the assessments are conducive to their impact on teaching and our value systems in general, the findings from this study can contribute the following to the validity of CWA.

7.3.3.1 Making Inferences

Being able to make inferences about a student's writing abilities based on the test scores relates to the relationship between the scores and the writing construct. The writing construct is reflected in the rubric (see Appendix B). Although the raters in this study attended to relevant features of the scripts in the communal ratings, they did not directly consult the scale descriptors of the rubric to guide them in assigning their scores. This may be perceived as negatively affecting the validity of the assessments, the argument being that if the raters do not consult the

scale descriptors, their assessments do not reflect the writing construct. Others might argue that it is precisely because the raters rely on human responses and contextualized judgments rather than the abstract scale descriptors that their assessments become valid. The core of this debate is, of course, how much faith one has in the ability of the scoring rubric to represent the writing construct in question.

7.3.3.2 Consequential Validity

Consequential validity refers to the potential impact, consequences and value implications of a test and its uses in a social context (see Section 2.3.4.2 above). This research did not analyze the uses made of the scores from the HHX in education, students' lives, or society in Denmark, so little can be said about these fundamental aspects of the test's consequential validity. However, findings from the research do point toward the potential long-term value that CWA may have for the raters in their development of assessments, their teaching practices as well as the underlying value implications.

7.3.3 2.1 Value Implications

Some tentative conclusions can be drawn on the epistemological, value and ethical underpinnings of CWA.

It can be argued that the tendency among CWA raters to not directly consult the scale descriptors when assigning scores can have some unethical consequences for the stakeholders. The rubric, with its scale descriptors, is the only explicit policy statement of the writing construct that the public has access to, and if raters do not refer to the scale descriptors directly, it will be difficult for people outside the rating community to infer from the test scores the level of writing ability that the students possess. What is comforting, however, is that compared to independent ratings communal ratings produce a more balanced attention to the textual features corresponding to the official assessment criteria in the rubric.

With respect to the epistemological and value implications of CWA suggested by the findings from the present study, CWA appears to enact a postmodern assessment practice. Broad (2000, 2003) and Huot (1996, 2002) have argued that postmodern writing assessment practices do not assume the existence of an absolute textual value. Such epistemological underpinnings are reflected in the findings of the present study by the raters' inclinations to not rely on the abstract scale descriptors directly but rather to rely on the inter-subjectivity of the assessments evolving from the communal rating discussions.

The CWA practice manifested in the present study also displays what Broad (2000) would refer to as a democratic assessment practice. Because CWA lets raters rely on the expertise of each other rather than subjecting them to absolute values determined by a rubric, the values and opinions of these raters (who are also teachers) are respected and heard. Minority voices as well as majority voices are considered ... whether they end up counting towards the final score or not. The end result (i.e. the final score) of a CWA session is sometimes a compromise between two discrepant scores, and sometimes it results from one rater conceding to the other rater. Concessions, however, appear to be made largely on the basis of careful deliberation and equal participation rather than dominance, although, of course, the complexity of dominance and its effects are difficult to determine.

7.3.3.2.2 Impact

The impact that CWA practices are likely to have on future assessments and teaching practices can only be speculated on here. However, the raters' decision-making behaviors in the communal rating sessions as well as their retrospective perceptions have given us some idea of the potential impact of CWA.

7.3.3.2.2.1 Impact on Future Assessment Practices

The opportunities given to the raters to validate their assessments and test out their insecurities against other raters pave the way for continuous rater development. The raters in this study

seemed to seize this opportunity because they voiced their insecurities and defined or revised their assessment strategies. Also, the fact that the raters generally perceived the communally rated scores to be more accurate than their own scores from the independent ratings points to the raters' tendency to learn from such CWA practices. In fact, the raters themselves explicitly referred to this benefit, or even necessity, when asked to reflect on CWA in general.

The decrease in extreme scores from independent ratings to communal ratings in this study points towards enhanced reliability potentials of CWA. As the raters are given a forum in which they can validate their assessment strategies and get an insight into the decision-making behaviors of other raters, they can gradually align their assessments. The development toward a progressively more balanced attention to the different textual features when moving from interpretation strategies over judgment strategies to strategies employed in the communal ratings indicates a tendency with the raters to monitor themselves into the assessment norms of the rater community.

7.3.3.2.2.2 Impact on Future Teaching Practices

CWA seems to value individual, multi-perspectival reader responses and contextualized assessments (e.g. Moss, 1994, 1996: Broad, 2000, 2003), and we can assume that such values will be reflected in writing instructions and learning activities in the classroom. As the raters in this study were all teachers teaching courses leading up to the specific test, their insight into how other raters/teachers/readers respond to different scripts can be valuable not only for themselves as teachers but also for the students who receive instruction from them. During the communal rating sessions the raters referred to their own teaching practices, which indicates some potential benefit for CWA to impact synergistically on teaching. As Broad (2003) pointed out, CWA procedures hold potential for the continuous development of teaching and assessment programs.

7.4 Implications for the Concept of Rater Expertise

Although the concept of rater expertise has not been the central focus of this study, some implications can be drawn on this concept both with respect to their behaviors in independent rating sessions and their behaviors in the communal rating sessions.

When rating independently, expert CWA raters might be characterized slightly differently than expert raters rating independently in psychometrically driven assessment practices. Wolfe et al. (1998) characterized proficient raters of such assessment practices as employing fast, rubric-adhering, principle-driven assessment processes. The highly experienced raters in this study relied on their co-raters rather than on the scale descriptors, made specific, rather than general references to the scripts, often exemplifying directly from them, and in general made thorough and thoughtful assessments. So it seems these CWA raters cannot be evaluated based on the same superficial and rubric-driven principles as raters in psychometrically driven practices.

As CWA builds on hermeneutic principles of validation and interactions between or among raters, one would assume that when rating communally expert CWA raters would be characterized by their ability to engage equally and critically with other raters so that variations can be brought out into the open and enabling and disabling prejudices be determined. This is exactly what the experienced CWA raters did in the present study. Further, the experienced raters in this study brought forth a balanced reference to different textual features of the student scripts and did not like the raters in Broad's (2003) study, who were new to CWA, restrict themselves to mechanics because such aspects were simply easier to refer to. So, it seems that these experienced CWA raters have learnt to express their assessments about less tangible textual features such as content and use of source materials.

7.5 Summary of Discussions of Findings

The CWA practice investigated in the present study showed that CWA raters seem to go through a complex rating process in which they conscientiously interpret and judge student scripts in their independent rating sessions to obtain preliminary scores for each script in preparation for the succeeding communal rating sessions, in which they carefully validate and refine their

assessments to assign what they believe to be the most accurate scores. For the most part, the results from the present study confirm the complex decision-making behaviors identified in traditional rating sessions in which raters rate scripts independently, and they support the theoretical claims for CWA and confirm the equal and critical engagement found in the few empirical studies conducted in communal ratings. The results, however, do not echo how the raters in other studies into independent ratings used their scoring rubric. Although the raters in the present study balanced their attention to the textual features corresponding to the official assessment criteria, they did not directly consult the scale descriptors, not even as a resolution or articulation tool. Instead they would rely on their co-raters for solving problems and resolving insecurities related to scoring. Further, the experienced raters in the present study did not replicate the frustrations experienced by the novice CWA raters in other studies. Experience with CWA seemed to have taught the raters that rater variation is inevitable and sometimes even desirable and fruitful, not only for score accuracy, but also for the raters' own professional development.

The validity potentials of CWA as demonstrated in the findings of the present thesis study seem promising, at least from a hermeneutic point of view: CWA appears to offer sound opportunities for raters to conduct thorough, well-deliberated, and multiple-perspectival assessments progressively leading to refined judgments. Further, CWA seems to reflect a postmodern epistemology and democratic values, it provides raters with ample opportunity to develop themselves as raters and teachers, and it paves the ground for improved teaching and assessment practices in general. From a psychometric point of view the progressively balanced attention to the textual features corresponding to the criteria prescribed in the scoring rubric seem promising. Less promising, however, would be the little difference in levels of agreement among raters and the tendency among CWA raters to rely on the rating community rather than the scale descriptors for score resolution.

With respect to the concept of rater expertise, expert CWA raters may be different from expert raters in traditional writing assessment practices. Unlike expert raters in traditional performance-based writing assessment, expert CWA raters may be characterized not by making general

references to the scripts and adhering closely to the scale descriptors, but rather by being able to perform thorough judgments, to point directly to strengths and weaknesses of student scripts, to engage critically and equally with other raters and accept differences as a synergetic potential in writing assessment.

7.6 Conclusion

This study has contributed to the small body of validation studies that exists in communal writing assessment (CWA) by mapping CWA raters' decision-making behaviors and chronicling the progression of these behaviors all the way from the raters' independent ratings to their communal ratings. The study was motivated by an increasing theoretical interest in CWA in the writing assessment literature and by a long and largely un-researched tradition of CWA practice in the Danish educational context.

Scholars within the fields of composition and testing have emphasized that, at least from a hermeneutic point of view, CWA offers increased validity potentials because raters in a CWA practice are given the opportunity to validate and refine their assessments during the course of the rating process thereby potentially generating increasingly sounder judgments of student scripts. Although some empirical studies have been conducted to support these claims, they are few, and to my knowledge, no study has looked systematically into CWA raters' rating process by tracing their decision-making behaviors from their independent rating sessions, where they form their preliminary scores, to their communal rating sessions, where they work with other raters to finalize their scores. This study did that in an exploratory way relying on descriptive, introspective, and discourse analytic research methods for a small number of experienced CWA raters during the administration of one CWA exam in Denmark.

The CWA in Denmark especially calls for empirical investigation because despite having been practiced for decades in the assessment of writing skills in the mother tongue as well as in English as a foreign language (even before the increased interest in such a procedure evolved in the international arena), it has not been subjected to systematical empirical analysis.

The present study confirms the sparse empirical research into CWA, which has shown that CWA fosters sound and contextualized judgments along with potentials for professional development. By systematically tracing raters' decision-making behaviors from their independent ratings to their communal ratings, the study has also shown that raters develop and refine their assessment strategies during the rating process. From a psychometric point of view, however, CWA might seem less promising as in such an assessment practice there does not seem to be much increase in reliability levels, and the raters largely ignore the scale descriptors of the rubric, preferring to rely on their co-raters rather than the scale. So, whether CWA can be considered a more valid than a "non-communal" writing assessment procedure largely depends on whether one takes a hermeneutic or a psychometric perspective and whether one believes that it is possible to construct a scoring scale that can adequately represent the writing construct.

As what has been reported on here is an exploratory case study, further studies are needed to assess the validity of CWA. Besides replicate studies to support the results of the present study, studies that focus on concurrent and predictive validation would bring to light a broader aspect of the validity potentials of CWA. Of particular interest, though, would be longitudinal studies that can inform us of the impact of CWA. As one of the claims to the validity of CWA is that raters can develop professionally over time, studies that look into how CWA raters develop their decision-making behaviors over the course of their CWA careers would cast further light on the developmental potentials inherent in CWA practices.

Summary in Danish – Resume på dansk

Indledning

Denne Ph.d. afhandling med titlen From Independent Ratings to Communal Ratings: A Study of CWA raters' Decision-Making Behaviors er en undersøgelse af, hvorledes censorer bedømmer elevers skriftlige færdigheder i en konsensusbedømmelsesform, hvor censorerne først bedømmer elevpræstationerne individuelt og derefter i en fællesbedømmelse med en anden censor skal nå til enighed om en endelig karakter.

Undersøgelsen har haft til formål at kortlægge progressionen af censorers bedømmelsesstrategier fra de individuelle bedømmelser til fællesbedømmelserne og diskutere konsensusbedømmelsesformens potentielle validitet.

Baggrund og Motivation

Hvor konsensusbedømmelsesformen betragtes som en innovativ bedømmelsesform i internationale sammenhænge, har den i flere årtier været en integreret del af det danske uddannelsessystem. Selvom man i andre lande undertiden inddrager mere end een censor, får censorerne sjældent mulighed for at konferere med hinanden, idet formålet med flere bedømmelser der er at øge reliabilitetskvotienten. I Danmark har man således i højere grad satset på (konstrukt)validitet, idet censorerne i konsensusbedømmelser får mulighed for at validere og forbedre deres bedømmelser af elevpræstationer sammen med andre censorer, før en endelig karakter afgives. Selvom vi i Danmark har betragtet konsensusbedømmelsesformen som en naturlig del af vores prøveformer, har den ikke tiltrukket megen forskningsinteresse, og vi har dermed ikke belæg for, om denne bedømmelsesform er valid og reliabel.

I udlandet (især Nordamerika) er nogle forskere dog begyndt at rette opmærksomheden mod konsensusbedømmelsesformen. En del teoretikere inden for såvel testning som skrivning har argumenteret for fordelene ved konsensusbedømmelser i forbindelse med bedømmelse af skriftlig udtryksfærdighed. Argumenterne grunder sig især på socialkonstruktivisme og en hermeneutisk tilgang til testning, idet der lægges vægt på intersubjektivitet og censorernes mulighed for sammen med andre at validere og dermed forbedre deres fortolkninger og bedømmelser af elevers skriftlige færdigheder.

Selvom det på det teoretiske plan ser ud til at konsensusbedømmelser kan bidrage til en (hermeneutisk) mere valid bedømmelse af skriftlige færdigheder, har der ikke været meget empirisk forskning inden for dette område. De få empiriske undersøgelser, der er foretaget, peger på, at konsensusbedømmelser også i praksis udviser et godt validitetspotentiale. Empirien er dog sparsom, og ingen har, så vidt jeg ved, foretaget en systematisk validitetsundersøgelse af konsensusbedømmelsesformen i sin helhed, dvs. fra censorerne bedømmer individuelt, til de sammen med en medcensor når frem til en endelig karakter.

Hovedformål og forskningsspørgsmål

Denne afhandling forsøger at råde bod på den sparsomme forskning i konsensusbedømmelser, idet den følger progressionen af censorers bedømmelsesstrategier i hele bedømmelsesproceduren, dvs. både i censorernes individuelle bedømmelser og i deres fællesbedømmelser. Fokus er primært på bedømmelsesprocessen, og sekundært på bedømmelsesproduktet (den endelige karakter). De mere specifikke forskningsspørgsmål, der udsprang af dette forskningsfokus var:

1. Hvilke specifikke bedømmelsesstrategier anvender censorerne i de individuelle bedømmelser og i fællesbedømmelserne, og hvad er sekvensen af disse bedømmelsesstrategier?
2. Hvordan fordeler censorerne deres opmærksomhed på de officielle bedømmelseskriterier i de individuelle bedømmelser og i fællesbedømmelserne?
3. I hvilken udstrækning findes der dominans i censordiskussionerne?

4. Hvad er karakterspredningen, og hvor stor er overensstemmelsen (konkordansen) mellem bedømmelserne i de individuelle bedømmelser og i fællesbedømmelserne?
5. Hvorledes vurderer censorerne konsensusbedømmelsesformen generelt og i forhold til de bedømmelser, de netop har deltaget i?

Metode

For at kunne besvare disse forskningsspørgsmål foretog jeg en empirisk undersøgelse af 20 censorers bedømmelsesstrategier i forbindelse med deres bedømmelse af 15 HHX stile skrevet på engelsk af danske HHX kandidater. De primære instrumenter til at kortlægge deres bedømmelsesstrategier var verbale protokoller: højtænkningsprotokoller blev anvendt i forbindelse med censorernes bedømmelsesstrategier i deres individuelle bedømmelser, og audiooptagelser anvendtes i fællesbedømmelserne. Retrospektive rapporter over censorernes opfattelse af konsensusbedømmelsesformen generelt og i forbindelse med den konkrete bedømmelsessituation blev brugt til at kaste yderligere lys over deres bedømmelsesstrategier. Desuden blev censorernes karakterfordeling undersøgt.

Resultater

Resultaterne er opsummeret nedenfor i forhold til de enkelte forskningsspørgsmål.

1. **Hvilke specifikke bedømmelsesstrategier anvender censorerne i de individuelle bedømmelser og i fællesbedømmelserne, og hvad er sekvensen af disse bedømmelsesstrategier?**

Censorerne i min undersøgelse foretog en grundig fortolkning og bedømmelse af de enkelte elevbesvarelser både i de individuelle bedømmelser og i fællesbedømmelserne. Når censorerne bedømte besvarelserne individuelt, startede de med at scanne dem for at få et overblik over længde, overskrift, og til dels den overordnede struktur. Dernæst gennemlæste de dem intensivt og dannede sig et billede af dem ved at fokusere primært på elevernes formuleringsevne (især grammatiske fejl), men også til en vis grad på, hvordan eleverne

håndterede indholdet og benyttede det læsemateriale, de havde haft til rådighed. For at komme frem til en bedømmelse fokuserede censorerne også meget på elevernes formuleringsevne, men forsøgte at opveje det fokus ved også at rette opmærksomheden på andre tekstuelle faktorer. Kontekstuelle faktorer som f.eks. opgaveformuleringen, den tid eleverne havde til rådighed under eksamen og sammenligning med andre elevbesvarelser blev også inddraget. Når censorerne var i tvivl om deres bedømmelser (både mht hvilken karakter der skulle gives, men også mht hvor meget enkelte punkter, så som misforståelser, for mange kongruensfejl, længde, osv. skulle tælle), konsulterede de ikke bedømmelsesvejledningerne eller den dertilhørende bedømmelsesskala, men foretrak at give fleksible karakterer og sætte deres lid til de efterfølgende censordiskussioner i fællesbedømmelserne.

I fællesbedømmelserne validerede censorerne deres bedømmelser, hvis der var uoverensstemmelse mellem deres indledende karakterer fra de individuelle bedømmelser. Ligesom i de individuelle bedømmelser blev der her lagt vægt på såvel tekstuelle som kontekstuelle aspekter af elevbesvarelserne, men i modsætning til de individuelle bedømmelser overskyggede fokus på elevernes formuleringsevne her ikke fokus på andre tekstuelle aspekter. Censorerne refererede ofte direkte til elevbesvarelserne og benyttede sig undertiden af muligheden for at diskutere deres bedømmelsesstrategier (f.eks. hvor meget en for kort besvarelse skulle trække ned). Hvis der var tvivl om en endelig karakter, konsulterede censorerne heller ikke her bedømmelsesvejledningen direkte, men revurderede snarere deres argumenter, sammenlignede med tidligere karaktergivne besvarelser eller lod den endelige karakter falde midt mellem deres individuelle karakterer.

2. Hvordan fordeler censorerne deres opmærksomhed på de officielle bedømmelseskriterier i de individuelle bedømmelser og i fællesbedømmelserne?

De officielle bedømmelseskriterier for HHX: Fylde, Strukturering, Anvendelse af tekstmaterialet og forståelse af problemstillingen, Formuleringsevne (ordforråd, syntaks, variation, idiomatik, grammatik) og Formalia blev taget i betragtning i såvel de individuelle

bedømmelser som i fællesbedømmelserne. Men hvor opmærksomheden på elevernes Formuleringsevne (især grammatik) overskyggede opmærksomheden på de andre bedømmelseskriterier i de individuelle bedømmelser, blev denne opmærksomhed på Formuleringsevne reduceret, når censorerne bedømte sammen, hvilket førte til en mere balanceret opmærksomhed på de opstillede bedømmelseskriterier.

3. I hvilken udstrækning findes der dominans i censordiskussionerne?

Det blev undersøgt, hvorvidt censorerne deltog på lige fod i karakterforhandlingerne i fællesbedømmelserne, eller om der var tale om dominans/undertrykkelse, både med hensyn til hvem der styrede samtalen (samtaledominans) og med hensyn til, hvis individuelle karakter lå nærmest den endelige karakter (karakterdominans). Det viste sig, at få censorer havde tendens til at dominere samtalen, og få censorer viste en tendens til at dominere i karaktergivningen. Men selvom der undertiden var tegn på dominans, lod karakterdominans ikke til at være et produkt af samtaledominans. Med andre ord, hvis en censor havde tendens til at dominere samtalen, havde denne censor ikke nødvendigvis tendens til at dominere karaktergivningen. At dominans ikke var et fremherskende fænomen i censordiskussionerne blev understreget i censorernes retrospektive kommentarer. I de tilfælde, hvor deres individuelle karakter var forskellig fra den endelige karakter, blev de spurgt hvilken af disse to karakterer, de mente, var den mest passende, og i langt de fleste tilfælde (95% af tilfældene) mente censorerne, at den endelige karakter var den mest passende karakter.

4. Hvad er karakterspredningen, og hvor stor er overensstemmelsen (konkordansen) mellem bedømmelserne i de individuelle bedømmelser og i fællesbedømmelserne?

Overensstemmelsen mellem bedømmelserne ændrede sig ikke fra de individuelle bedømmelser til fællesbedømmelserne, men der viste sig i nogle tilfælde at være forskel i karakterspredningen. Hvor der i de individuelle bedømmelser kunne være op til 4 karakterers forskel, var der sjældent mere end 1 eller 2 karakterers forskel i fællesbedømmelserne. At konsensusbedømmelsesformen var med til at eliminere de mest ekstreme karakterer blev underbygget i censorernes retrospektive kommentarer, hvor det bl.a. blev nævnt, at en af

fordelene ved konsensusbedømmelsesformen er, at den er med til at mindske betydningen af idiosynkrasier og fejlbedømmelser.

5. Hvorledes vurderer censorerne konsensusbedømmelsesformen generelt og i forhold til de bedømmelser de netop har deltaget i?

Alle censorer blev spurgt retrospektivt om deres vurdering af konsensusbedømmelsesformen, både i forhold til konsensusbedømmelser generelt og i forhold til de bedømmelser, de netop havde deltaget i. Som nævnt ovenfor var censorerne stort set enige om, at de konsensusbedømmelser, som de netop havde deltaget i, havde ført til mere præcise karakterer. Denne positive opfattelse af konsensusbedømmelserne var reflekteret i deres opfattelse af konsensusbedømmelsesformen generelt. Censorernes udelukkende positive kommentarer gik på at konsensusbedømmelser generelt producerer mere præcise karakterer, at de er med til at opretholde en national standard, og at de giver censorerne mulighed for at udvikle sig professionelt.

Konklusion

Afhandlingens overordnede konklusion er, at konsensusbedømmelser af elevers skriftlige færdigheder potentielt kan føre til mere valide bedømmelser, især fra et hermeneutisk synspunkt.

Mine resultater bekræfter tidligere forskning i konsensusbedømmelser, der har vist at censorer engagerer sig kritisk og på lige fod med hinanden for at finde frem til en endelig karakter. Derudover viser min undersøgelse, at der sker en udvikling i censorernes bedømmelsesstrategier fra de individuelle bedømmelser til fællesbedømmelserne: censorerne retter deres opmærksomhed mere ligeligt på de tekstuelle faktorer, der svarer til de opstillede bedømmelseskriterer. Samtidig mener censorerne selv, at fællesbedømmelserne afstedkommer mere præcise karakterer end de individuelle bedømmelser.

Til trods for at ekstreme karakterer ser ud til at kunne elimineres i konsensusbedømmelser, vil denne bedømmelsesform ud fra et psykometrisk perspektiv dog ikke kunne bidrage meget til et

øget validitetspotentiale, først og fremmest fordi overensstemmelsen mellem karaktererne ikke øges fra de individuelle bedømmelser til fællesbedømmelserne, men også fordi censorerne ikke direkte konsulterer bedømmelsesvejledningerne, især bedømmelsesskalaerne.

Det at censorer ikke konsulterer bedømmelsesvejledningerne er heller ikke et ukendt fænomen i bedømmelsesprocedurer, der ikke anvender konsensusbedømmelser. Flere undersøgelser i censorers bedømmelsesprocesser har vist, at censorer i sådanne mere psykometrisk-fokuserede bedømmelsesesformer ofte ikke konsulterer bedømmelsesskalaerne, fordi de finder det svært at forene en abstrakt skala med deres varierede og komplekse fortolkninger og bedømmelser af elevtekster. Man kan således betragte konsensusbedømmelsesformen som specielt velegnet til bedømmelse af et tekstfag som skriftlig udtryksfærdighed, der på grund af stilistisk variation er en disciplin, der kræver fortolkning.

References

Allen, M. S. (1995): Valuing differences: Portnet's first year. *Assessing Writing 2, (1)*, 67-89.

Allwright, D. & Bailey, K. (1991). *Focus on the language classroom: An introduction to classroom research for language teachers*. Cambridge: Cambridge University Press.

Bachman, L. F. (1997): *Fundamental considerations in language testing.* Oxford and New York: Oxford University Press.

Bachman, L. F. (2004): *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Barkaoui, K. (2007a): *Effects of thinking aloud on ESL essay rater performance: A FACETS analysis*. Paper presented at the 29th Annual Language Testing Research Colloquium.

Barkaoui, K. (2007b): Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *The Canadian Modern Language Review*, 64(1), 99-134.

Belanoff, P. (1991): The myths of assessment. *Journal of Basic Writing, 10(1)*, 54-66.

Belanoff, P. & Denny, H. (2006): Review of the book [*What we really value: Beyond rubrics in teaching and assessing writing*]. *Journal of Teaching Writing, 22(2)*, 127-138.

Berger, P. & Luckmann, T. (1966): *The social construction of reality: A treatise in sociology of knowledge*. Garden City, NY: Doubleday.

Bracewell, R.J. & Breuleux, A. (1994): Substance and romance in analyzing think-aloud protocols. In Smagorinsky, P. (Ed.): *Speaking about writing: Reflections on research methodology* (pp. 55–88). Thousand Oaks, CA: Sage.

Breland, H. (1983): *The direct assessment of writing skill: A measurement review* (Technical Report No. 83-6). Princeton, NJ: College Entrance Examination Board.

Breland, H., Lee, Y, Najaran, M. & Muraki, E. (2004): An analysis of TOEFL CBT writing prompts difficulty and comparability for different gender groups. *TOEFL Research Report No. RR-74*. Princeton, NJ: Educational Testing Service.

Broad, R. L. (1994): "Portfolio scoring": A contradiction in terms. In Black, L., Daiker, D., Sommers, J. & Stygall, G. (Eds.): *New directions in portfolio scoring* (pp. 263-276). Portsmouth: Boynton/Cook.

Broad, B. (1997): Reciprocal authorities in communal writing assessment: Constructing textual value within a "new politics of inquiry". *Assessing Writing 4(2),* 133-167.

Broad (2000): Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English, 35(2)*, 213-260.

Broad, B. (2003): *What we really value: Beyond rubrics in teaching and assessing writing.* Logan, UT: Utah State University Press.

Broad, B. & Boyd, M. (2005): Rhetorical writing assessment: The practice and theory of complementarity. *Journal of Writing Assessment, 2(1),* 7-20.

Charney, D. (1984): The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18(1)*, 65-81.

Coffman, W. (1971): On the reliability of ratings of essay examinations in English. *Research in the Teaching of English, 5*, 24-36.

Condon, W. & Hamp-Lyons, L. (1994): Maintaining a portfolio-based writing assessment: Research that informs program development. In Black, L., Daiker, D.A., Sommers, J. & Stygall, G. (Eds.): *New directions in portfolio assessment* (pp. 277-285). Portsmouth, NH: Boynton/Cook.

Connor-Linton, J. (1995a): Cross-cultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes, 14(1),* 99-115.

Connor-Linton, J. (1995b): Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29(4)*, 762-765.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31-51.

Cumming, A., Kantor, R., & Powers, D. (2001): *Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision making, and development of a preliminary analytic framework.* TOEFL Monograph Series, Report No. 22. Princeton, NJ: Educational Testing Service.

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

Daly, J.A. & Dickson-Markman (1982): Contrast effects in evaluating essays. *Journal of Educational Measurement, 19(4)*, 309-316.

Davies, A., Brown, A., Elder, C. Hill, K. Lumley, T. & McNamara, T. (2002): *Dictionary of language testing, Studies in Language Testing 7.* Cambridge: Cambridge University Press.

Dechert, H.W. (1987). Analysing language processing through verbal protocols. In Faerch & Kasper (Eds.): *Introspection in second language research*, (pp. 96-112). Clevedon: Multilingual Matters.

Delandshere, G. & Petrosky, A. R. (1994): Capturing teachers' knowledge. *Educational Researchers, 23(5),* 11-18.

DeRemer, M.L. (1998): Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5(1),* 7-29.

Deville, G. & Chalhoub-Deville, M. (2006): Old and new thoughts on test score variability: implications for reliability and validity. In M. Chalhoub-Deville, C. A., Chapelle & P. Duff (Eds.), *Inference and generalizability in applied linguistics: multiple perspectives* (pp. 9-25). Amsterdam: Benjamins.

Diederich, P. B., French, J. W, & Carlton, F. S. (1961): *Factors in judgments of writing quality.* Princeton, NJ: Educational Testing Service: RB No. 61-15. ED 002 172.

Durst, R. Roemer, M. & Schultz, L. (1994): Portfolio negotiations: Acts in speech. In L. Black, D. A. Daiker, J. Sommers, & G. Stygall (Eds.): *New directions in portfolio assessment* (pp. 286-300). Portsmouth, NH: Boynton/Cook.

Ecke, T. (2008): Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155-185.

Elbow, P. (1991): Foreword. In P. Belanoff & M. Dickson (Eds.): Portfolios: Process and product. Portsmouth, NH: Boynton/Cook.

Elbow, P. (1993): Ranking, evaluating, and liking: Sorting out three forms of judgment. *College English 55(2)*, 187-206.

Elliot, N. (2005): *On a scale: A social history of writing assessment in America.* New York: Peter Lang Publishing Inc.

Erdosy, M. U. (2004): *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions.* Report 70. Princeton, NJ: Educational Testing Service.

Ericsson, K. A. & Simon, H. A. (1984/1993): *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Evalueringsinstitut (2005a): *Censorinstitutionen – status 2005.* Copenhagen: Danmarks Evaluerings Institut. www.eva.dk

Evalueringsinstitut (2005b): *Censorundersøgelse og rapport om taxametersystemet og uddannelseskvalitet.* Copenhagen: Danmarks Evalueringsinstitut. www.eva.dk.

Freedman, S. W. (1979): How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71(3)*, 328-338.

Freedman, S. W., & Calfee, R.C. (1983): Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S.A. Walmsley (Eds.): *Research on Writing: Principles and Methods* (pp. 75-98). New York: Longman.

Gadamer, H. G. (1994): *Truth and method.* New York: Seabury.

Glaser, B. G. (1978): *Theoretical sensitivity: Advances in the methodology of grounded theory.* Mill Valley, CA: Sociology Press.

Green, A. (1998): *Verbal protocol analysis in language testing research. Studies in Language Testing 5*. Cambridge: Cambridge University Press.

Grobe, C. (1981): Syntactic maturity, mechanics and vocabulary as predictors of quality ratings. *Research in the Teaching of English, 15(1),* 75-86.

Grotjahn, R. (1987): On the methodological basis of introspective methods. In C. Faerch & G. Kasper (Eds.): *Introspection in second language research*, (pp. 54-81). Clevedon: Multilingual Matters.

Guba, E. G., ed. (1990): *The paradigm dialogue*. Newbury Park, CA: Sage.

Habermas, J. (1996): *Between facts and norms: Contributions to a discourse theory of law and democracy* (trans. William Rehg). Cambridge, MA: MIT Press.

Hare, A. P. (1976): *Handbook of small group research* (2nd Ed.). New York: Free Press.

Haue, H. (2000): Prøver og eksamen – norm og udfordring – set i et historisk perspektiv. *Uddannelse*, 4. Copenhagen: Undervisningsministeriet.

Hayes, J. R., & Flower, L. (1983): Uncovering cognitive processes in writing: An introduction to protocol analysis. In Mosenthal, P., Tamor, L. & Walmsley, S. (Eds.): *Research in writing: Principles and methods,* (pp.206-229). New York: Longman.

Heller, J. I., Sheingold, K. & Myford, C. M. (1998): Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5(1)*, 5-40.

http://eng.uvm.dk/factsheets/quality.htm?menuid=2505

http://eng.uvm.dk/factsheets/?menuid=25

http://eng.uvm.dk/publications/factsheets/htx08o_000.htm?menuid=2515

http://us.uvm.dk/grundskole/generelinformation/vejlendendetimetal/timetal/pdf

Homburg, T. J. (1984): Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18 (1),* 87-107.

Hoy, D. (1994): Critical theory and critical history. In Hoy, D. & McCarthy, T.: *Critical Theory* (pp. 101-214 & 249-273). Oxford: Blackwell.

Huot, B. (1990): The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60(2)*, 237-263.

Huot, B. A. (1993): The influence of holistic scoring procedures on reading and rating student essays. In Williamson, M. M. & Huot, B. A. (Eds.). *Validating holistic scoring for writing assessment* (pp. 206-236). Hampton Press, NJ: Inc. Cresskill.

Huot, B. (1996): Toward a new theory of writing assessment. *College Composition and Communication, 47(4),* 549-566.

Huot, B. (2002): *(Re)articulating writing assessment for teaching and learning.* Logan, UT: Utah State University Press.

Huot, B. & Schendel, E. (2001): Reflecting on assessment: Validity inquiry as ethical inquiry. *Journal of Teaching Writing, 17*, 37-55.

Itakura, H. (2001): *Conversational dominance and gender.* Amsterdam: John Benjamin's publishing Company.

Johnson, R.L., Penny, J., Gordon, B., Shumate, S.R., & Fisher, S.P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly, 2(2)*, 117-146.

Johnston, P. (1989): Constructive evaluation and the improvement of teaching and learning. *Teachers College Record, 90(4)*, 509-28.

Kobayashi, T. (2002): Native and nonnative reactions to ESL compositions. *TESOL Quarterly*, 26, 81-112.

Kobayashi, H. & Rinnert, C. (1996): Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning, 46(3)*, 397-437.

Kondo-Brown, K. (2002): A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19, 3-31.

Kroll, B. (1998): Assessing writing abilities. *Annual Review of Applied Linguistics, 18*, 219-240.

Lee, Y., Breland, H. & Muraki, E. (2004): Comparability of TOEFL CBT writing prompts for different native language groups. *TOEFL Research Report, No. RR-77.* Princeton, NJ: Educational Testing Service.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing 19(3),* 246-276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective.* Frankfurt am Main: Peter Lang.

Lynne, P. (2004): *Coming to terms: Theorizing writing assessment in composition studies.* Logan, UT: Utah State University.

Matsumoto, K. (1993): Verbal-report data and introspective method in second language research. *RELC Journal*, 24(1), 32-60.

Mendelsohn, D. & Cumming, A. (1987): Professors' ratings of language use and rhetorical organizations in ESL compositions. *TESL Canada Journal/Revue TESL du Canada, 5(1)*, 9-26.

Messick, S. (1989): Validity. In Linn, R. L. (Ed.) *Educational Measurement,* pp. 13-104. (3rd edition). New York: Macmillan

Milanovic, M., Saville, N. & Shuhong, S. (1996): A study of the decision-making behavior of composition markers. In Milanovic, M. & Saville, N. (Eds.): *Performance testing, cognition and assessment. Selected Papers from Studies in Language Testing 3*, 92-114. Cambridge: Cambridge University Press.

Miles, M.B., & Huberman, A.M. (1984). *Qualitative data analysis*. Newbury Park, CA: Sage.

Miles, M.B, & Huberman, A.M. (1994). *Qualitative data analysis* (2nd Edition). Newbury Park, CA: Sage

Mislevy, R. (1994): *Can there be reliability without "reliability?"* Princeton, NJ: Educational Testing Service, 1-24.

Mohan, B. & Low, M. (1995): Collaborative teacher assessment of ESL writers: Conceptual and practical issues. *TESOL Journal, Autumn*, 28-31.

Moss, P. (1994): Can there be validity without reliability? *Educational Researcher, 23(2)*, 5-12.

Moss, P. (1996): Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25(1)*, 20-29.

Moss, P., Schutz, A, & Collins, K. (1998): An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education, 12(2)*, 139-161.

Moss, P. & Schutz, A. (2001): Educational standards, assessment, and the search for consensus. *American Educational Research Journal, 28(1)*, 37-70.

Myers, M. (1980): *A procedure for writing assessment and holistic scoring.* Urbana, IL: NCTE.

Nixon, R. & McClay, J. K. (2007): Collaborative writing assessment: Sowing seeds for transformational adult learning. *Assessing Writing, 12*, 149-166.

O'Laughlin, K. (1994): The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics,* 17, 23-44.

Paltridge, B. (1994): Genre analysis and the identification of textual boundaries. *Applied Linguistics, 15(3)*, 288-299.

Penny, J. Johnson, R. L., & Gordon, B. (2000): Using rating augmentation to expand the scale of an analytic rubric. *The Journal of Experimental Education, 68(3),* 269-287.

Perl, S. (1979): The composing processes of unskilled college writers. *Research in the Teaching of English, 13(4)*, 317-36.

Petersen, N. R. (2001). *Elementær statistik* (2nd Edition). Copenhagen: Institut for Almen og Anvendt Sprogvidenskab.

Petersen, P. B. (2006): Fra rosværdig til 13. *Uddannelse*, 1. Copenhagen: Undervisningsministeriet.

Plischewski, N. (2003): Evaluering, prøver og eksamen – kvalitetssikring og kvalitetsudvikling. *Uddannelse*, 3. Copenhagen: Undervisningsministeriet.

Pressley, M. & Afflerbach, P. (1995): *Verbal protocols of reading: The nature of constructively responsive reading.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Polio, C. & Glew, M. (1996): ESL writing assessment prompts: How students choose. *Journal of Second Language Writing, 5, 35-49.*

Pula, J. J. & Huot, B. A. (1993). A model of background influences on holistic raters. In Williamson, M. M. & Huot, B. A. (Eds.). *Validating holistic scoring for writing assessment* (pp. 206-236). Hampton Press, NJ: Inc. Cresskill.

Raimes, A. (1985). What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly, 19,* 229-258.

Rinnert, C. & Kobayashi, H. (2001): Differing perceptions of EFL writing among readers in Japan. *Modern Language Journal*, 85, 189-209.

Sakyi, A.A. (2000): Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In Kunnan, A., J. (Ed.): *Fairness and validation in language assessment. Selected papers from the 19th Language Testing Colloquium, Orlando, Florida* (pp. 129-152). Cambridge, England: Cambridge University Press.

Sakyi, A.A. (2003): *A study of the holistic scoring behaviours of experienced and novice ESL instructors.* Unpublished Ph.D. thesis. Department of Curriculum Teaching and Learning. The Ontario Institute for Studies in Education of the University of Toronto.

Shi, L. (2001): Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18(3),* 303-325.

Sielemann, K. Chairman of the national rater corps for HHX, English.

Smagorinsky, P. (Ed.) (1994a): Introduction: potential problems and problematic potentials of using talk about writing as data about writing process. In Smagorinsky, P. (Ed.): *Speaking about writing: Reflections on research methodology (*pp. ix-xix*).* Thousand Oaks, CA: Sage.

Smagorinsky, P. (Ed.) (1994b): Think-aloud protocol analysis: Beyond the black box. In Smagorinsky, P. (Ed.): *Speaking about writing: Reflections on research methodology (pp. 3-19).* Thousand Oaks, CA: Sage.

Song, B. & Caruso, I. (1996): Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5(2)*, 163-182.

Spolsky, B. (1995): *Measured words.* Oxford: Oxford University Press.

Stansfield, C.W. & Ross, J. (1988): A long-term research agenda for the test of written English. *Language Testing*, 5(2), 160-186.

Stewart, M.R. & Grobe, C.H. (1979). Syntactic maturity, mechanics, and vocabulary and teachers' quality ratings. *Research in the Teaching of English, 13*, 207-215.

Stock, P.L. & Robinson, J. L (1987): Taking on testing: Teachers as testers researchers. *English Education, 19(2)*, 93-121.

Stratman, J.F. & L. Hamp-Lyons. (1994): Reactivity in concurrent think-aloud protocols: Issues for research. In Smagorinsky, P. (ed.): *Speaking about writing: Reflections on research methodology* (pp. 89-114). Thousand Oaks, CA: Sage.

Swarts, H., Flower, K., & Hayes, J.R. (1984): Designing protocol studies of the writing process. In R. Beach and L. Bridwell (Eds.): *New directions in composition research,* (pp. 43-71). New York: Guilford.

Sweedler-Brown, C.O. (1985): The influence of training and experience on holistic essay evaluation. *English Journal*, 74, 49-55.

Sweedler-Brown, C. O. (1993): ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing*, 2(1), 3-17.

Uddannelsesstyrelsens internetpublikationer (2001): Råd og vink om hhx generelt (nr. 8). Chapter 7: Årsprøver og eksamen. Copenhagen: Undervisningsministeriet. http://us.uvm.dk/gymnasie/erhverv/generelt/raadogvink/haeftegenerelthhx/indhold.htm

Undervisningsministeriet (2004): Betænkning om indførelse af en ny karakterskala til erstatning af 13-skalaen. Copenhagen: Undervisningsministeriet. http://pub.uvm.dk/2004/karakterer

Vann, R.J., Lorenz, F.O. & Meyer, D. M. (1991): Error gravity: Faculty response to errors in the written discourse of nonnative speakers of English. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 181-195). Norwood, NJ: Ablex.

Vaughan, C. (1991): Holistic assessment: What goes on in the rater's mind? In Hamp-Lyons, L. (Ed.): *Assessing Second Language Writing in Academic Contexts* (pp.111-25). Norwood, NJ: Ablex.

Weigle, S. C. (1994): Effects of training on raters of ESL compositions. *Language Testing, 11(2)*, 197-223.

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

Wiggins, G. (1993): *Assessing student performance.* San Francisco, CA: Jossey Bass.

Wiggins, G. (1994): The constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing 1(1),* 129-39.

Williamson, M. M. (1993): An introduction to holistic scoring: The social, historical, and theoretical context for writing assessment. In Williamson, M. M. & Huot, B. A. (Eds.):

Validating holistic scoring for writing assessment: Theoretical and empirical foundations (pp. 1-43). Cresskill, NJ: Hampton.

Williamson, M. M. (1994): The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing, 1(2),* 147-74.

Wolfe, E. W. (1997): The relationship between essay reading style and scoring proficiency in a psychometric scoring System. *Assessing Writing, 4(1),* 83-106.

Wolfe, E., W., Kao, C.-W. & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15(4),* 465-492.

www.ealta.eu.org/documents/archive/guidelines/Englsih.pdf

www.iltaonline.com/code.pdf

www.ug.dk/vejledningsportal/Elementer/Guide%20til/Artikler.aspx?article_id=artik-hhxengelska

www.uvm.dk

www.us.uvm.dk/gymnasie/erhverv/eksamen/Vejledninger/vejl99en.html

www.utdanningsdirektoratet.no (Rammeverk for nasjonale prøver 2007)

Højere Handelseksamen
Handelsskolernes enkeltfagsprøve
maj 2004

04-0-2

Engelsk Niveau A

Grundtekst

Torsdag den 13. maj 2004
kl. 8.30-13.30

Grundtekst

Sewing a seam of worker democracy in China

Free trade union elections in foreign-owned factories are unprecedented in mainland China, says Alison Maitland. **Reebok's experiment may set a pattern for others**

A hush descends on the huge crowd of workers in the canteen at the Fu Luh factory in southern China that makes sports shoes for Reebok. Candidates in the first free elections to the factory's trade union are about to make their campaign speeches.

Many of the self-nominated candidates are shy. Some are so awed by the occasion that they cannot complete their speeches. Some read prepared texts. The bolder ones speak without notes. A brave few criticise the unelected, outgoing union officials for doing little to protect workers' rights and improve conditions. The audience, most of them young women workers from rural areas, listen intently.

When the voting slips from the secret ballot are counted in the factory courtyard, candidates who have spoken out for workers' interests emerge strongly represented among the 19 female and 12 male winners. The incumbent chairwoman, who is supported by local officials of the state-controlled union, has been voted out.

This experiment in worker democracy at the Taiwanese-owned factory in Fujian province two months ago, together with a similar election last year at a Reebok shoe manufacturer in neighbouring Guangdong, is thought to be unprecedented in mainland China.

"They are the first two foreign investment companies to have open union elections, with the brand company and factory support, as well as outside organisations as observers," says Monina Wong of the Hong Kong Christian Industrial Committee, a labour rights campaigning organisation that witnessed both elections.

[...]

The managers of the two factories arranged the elections at the behest of Reebok, their US client. "If Reebok wanted it and Reebok was pushing for it, they were going to go along with it," says Jonathan Unger, director of the Contemporary China Centre at the Australian National University (ANU). Mr Unger spent 13 days observing the Fu Luh election process and interviewing participants with his wife Anita Chan, a fellow China labour specialist, and two Chinese research assistants.

Reebok is now speaking publicly about the initiative for the first time. Last year's election at the Hong Kong-managed Kong Tai plant in Guangdong was the first of its kind for the company. "We were reluctant to do anything that might jeopardise its success," explains Doug Cahn, Reebok's director of human rights programmes.

[...]

At Fu Luh, it took months of negotiations between Reebok, the Taiwanese management and the official district trade union to agree a new constitution and election procedure. "It was no easy task," says Ms Chan. The complicated election used proportional representation to reflect the number of workers in each of the factory's seven departments. It went a step further than Kong Tai, where the office of union chairman was not contestable.

Reebok says its aim with these elections is to produce a sustained improvement in working conditions by promoting better communication between management and the shop floor. Multinational footwear and clothing brands find it notoriously difficult to ensure round-the-clock compliance with their codes of conduct on labour standards and human rights and are vulnerable to attack over abuses by their overseas suppliers.

"We have a code of conduct that says we will respect the rights of workers to freedom of association and collective bargaining," says Mr Cahn. "We can throw up our hands in China and say: 'The ACFTU is government-controlled and therefore we can do nothing.' Or we can engage in experiments like this in democratising the union in the hope that workers will take advantage of the opportunities this provides them."

The ACFTU: The All China Federation of Trade Unions

Some employers might consider it strange that Reebok wants to promote active unions in overseas factories on which it depends for its supplies. Mr Cahn argues that it is good for business. Better working conditions should strengthen the loyalty of the workforce, which in turn should help the management. He wants other multinationals to follow suit.

"In this part of our business, we don't seek to compete but to collaborate. Our ability to be successful in implementing a code of conduct is enhanced when there's a critical mass of multinational corporations or brands that are like-minded and are sending similar messages to the manufacturing community," he says.

This is not about gaining a competitive edge in the market, he insists. "I don't know that anybody has bought a pair of Reebok shoes because of its human rights programme. But we're a global corporation and we have an obligation to give back to the communities in which we live and work."

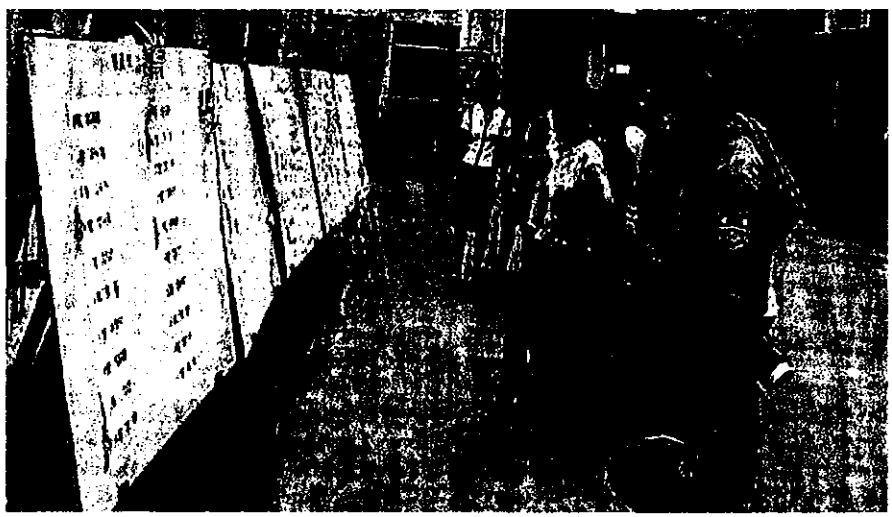
Mr Cahn acknowledges that Reebok would benefit from "a level playing field" with other big brands. It costs money to improve health and safety and make life more comfortable for workers. But he argues that there can be savings too, in reducing accidents and labour turnover.

Chief among the grievances that Chinese workers report to Reebok staff through confidential channels are abusive supervisors and excessive overtime hours. "It's our hope that issues can be taken up by the worker representatives," says Mr Cahn. "We have inspections of factories, both announced and unannounced. But you just don't have the assurance that things will be the same the next day. Factories in China are incredibly sophisticated at finding ways to fool us. The best monitors are the workers themselves."

Reebok has arranged training for the Kong Tai worker representatives with organisations such as aid agencies, to learn how to handle union matters such as conducting meetings and recording grievances. The Fu Luh representatives are due to have similar training.

Ms Chan and Mr Unger say the newly elected officials may be hampered by their inexperience, their lack of role models and the high workforce turnover. It may take time for them to make their voices heard. Nonetheless, they see the elections as a big step forward. Shortly after the elections at Fu Luh, workers were approaching the new committee for help, says Mr Unger. "Some of them are pretty strong people and they would go to management and say: 'Look, what you're doing here is illegal.' "

Both new unions are officially affiliated with the ACFTU. The challenge now for Reebok and the factory management, says Ms Wong, is to respect them for what they are: independent representatives of workers' interests.



Voting pattern: one aim of the elections to the trade union is to promote better communications between management and shop floor

Reebok fører det etiske korstog

Flere og flere danske virksomheder flytter produktionen til udlandet. Uden at have moralske skrupler over underbetaling, børnearbejde, diskrimination og andre brud på menneskerettighederne. Forbrugerombudsmanden vil nu bruge bl.a Reeboks etiske regler som forbillede for danske virksomheder. [Børsen] Fredag har besøgt Reebok, som også er betegnelsen for en sydafrikansk gazelle.

Af Susanne Tholstrup

Produktionen af Reebok fodbolde stoppede brat på en stor fabrik i Pakistan for få uger siden.
Kontrollanter og advokater fra Reeboks hovedkvarter i USA er i hast fløjet til Pakistan for at undersøge rygterne om, at virksomheden anvender børn som arbejdskraft på fabrikken.
Bekræfter eksperterne rygterne, er konsekvensen uomtvistelig: Samtlige Reebokbolde i fabrikken vil blive brændt, og produktionsaftalen med virksomheden opsagt med øjeblikkelig virkning. Mens undersøgelsen står på, forlader ingen bolde fabrikken.
Reebok accepterer ikke forretningspartnere, der benytter børn i produktionen, eller forretningspartnere, der benytter materialer lavet af børn. Den konsekvente handling, afbrænding af produktionen og opsigelse af samarbejdsaftalen, udløser Reebok også, hvis forretningspartnere underbetaler arbejderne, gennemtvinger arbejdstid, der fast ligger ud over 48 timer om ugen, benytter straffefanger i produktionen, eller diskriminerer ansatte på grund af hudfarve, religion, politisk observans eller national oprindelse.

Anstændighed

Der skal desuden være frihed til faglig organisering på enhver fabrik, der producerer sportssko, tøj og bolde for Reebok. Og der skal være vilje til på fabrikkerne at skabe et anstændigt miljø for de ansatte.
"Reebok Human Rights Production Standards" blev formuleret for fem år siden af den 100 år gamle virksomheds nuværende hovedaktionær, amerikaneren Paul Fireman, som i årtier har været aktiv i kampen for overholdelse af menneskerettighederne.
Reebok har ansat et korps af "revisorer" og advokater, hvis eneste opgave er at rejse rundt i verden og kontrollere, at virksomhedens etiske regler bliver overholdt i de cirka 40 fabrikker i Sydkorea, Indonesien, Kina, Thailand, Taiwan, Filippinerne, Italien og Spanien, som producerer Reebokprodukterne.
Nogle af virksomhederne beskæftiger op til 15.000 medarbejdere. I en detaljeret kontrakt med Reebok forpligter de sig til at overholde det etiske regelsæt, og i kontrakten er der aftalt en mindsteløn, der mindst skal svare til det niveau, som den lokale fagforening eller myndighed benytter. Men forudsat, at de ansatte kan leve anstændigt af lønnen, og at lønnen i den omkringliggende industri ikke er højere.

Den ægte vare

"Vores forpligtelser over for menneskerettighederne er et tiltag, som mange arbejdere er engagerede i. Det giver dem en stolthed, loyalitet og korpsånd i firmaet, som overgår alt, hvad vi kan gøre kommercielt", skriver Paul Fireman i et responsum.
Reeboks danske direktør, Steen Hummeluhr, er ikke i stand til at oplyse, hvor meget Reebok globalt spenderer på de etiske regler.
Beløbet er aldrig gjort op i koncernregnskabet. Menneskers vilkår ligger os blot voldsomt på sinde, og vi har aldrig udnyttet vores etik i markedsføringen eller ved at klistre et særlige logo på vore produkter. Vi tog de moralske standpunkter, længe før

der gik mode i etikken, siger Steen Hummeluhr. Han var derfor overrasket over forbrugerombudsmand Hagen Jørgensens skriftlige henvendelse til Reebok for nylig. Forbrugerombudsmanden bad om et eksemplar af Reeboks etiske regelsæt som inspiration til en etisk vejledning for danske virksomheder, som også i stigende omfang flytter produktionen ud i lavtlønsområderne i Asien og Østeuropa.

Baggrunden for Reeboks etiske principper blev skabt i 1988, da firmaet gik sammen med Amnesty International om at markere 40 året for UN Declaration of Human Rights med en koncert-turné med bl.a. Bruce Springsteen, Sting og Peter Gabriel.

Reebok-prisen

Samme år stiftede Reebok en pris, som hvert år uddeles til et ungt menneske, der har gjort en særlig indsats for menneskerettighederne.

Prisen på ca. 25.000 dollars er bl.a. givet til en af de kinesiske ledere af studenteroprøret på den Himmelske Freds Plads.

I 1992 gik prisen til den dengang 12-årige pakistanske dreng, Iqbal Masih, der som fire-årig blev solgt som slavearbejder til en tæppefabrik. I seks år arbejdede han 12-14 timer i døgnet. Da han slap ud af slaveriet, rejste han rundt i forskellige lande og fortalte sin historie, hvilket har skabt debat og boykot af flere lande, som benytter slave- og børnearbejde. I februar i år blev Iqbal Masih myrdet, 15 år gammel.

Paul Fireman er formand for Reeboks fond for menneskerettigheder.

Med i bestyrelsen er bl.a. Jimmy Carter, Sting og Peter Gabriel. I 1992 lancerede fonden programmet »Witness« med det formål at skænke videoudstyr, telefax, computere etc. til de personer og organisationer, som arbejder på at afsløre brud på menneskerettighederne rundt omkring i verden.

Direktør Steen Hummeluhr: Kampen for overholdelse af menneskerettighederne har kun mening, når der ligger konsekvent handling bag ordene. Derfor reagerer vi altid prompte på selv rygter om krænkelse af vores principper. Er rygterne falske, kræver vi dementi. Da formanden for Dansk Beklædnings- og Tekstilarbejderforbund for nogle år siden fremsatte den urigtige påstand, at kinesiske straffefanger producerede for Reebok, slap vi ikke sagen, før dementiet var ordentligt i hus, siger Steen Hummeluhr.

Hjemlig trivsel

Selv om overtrædelse af menneskerettighederne ikke ligefrem er et udbredt problem i Danmark, har også de 32 ansatte herhjemme glæde af koncernens etiske principper. Vi gør meget ud af medarbejdernes trivsel. Informationsniveauet i virksomheden er højt. Vi dyrker sport sammen og er ikke blege for at give en ekstra skilling til en kvindelig medarbejder, der skal på barselsorlov. Vi forsøger også at leve op til et samfundsmæssigt ansvar ved at skabe job til handicappede, langtidsledige eller folk på vej ned ad den sociale rangstige, siger Steen Hummeluhr.

Gazelle-spring

Reebok omsætter for 90 millioner kroner om året i Danmark. 250 modeller i sportssko sørger for 75 procent af omsætningen, mens en fjerdedel kommer fra sportstøj. Globalt omsætter Reebok for 3,5 milliarder dollars. Indsatsen for menneskerettighederne har ikke stået i vejen for væksten, som for alvor tog fart i 1980'erne, hvor Reebok, som er betegnelsen for en sydafrikansk gazelle, var den hurtigst voksende virksomhed i USA.

Fra 1984 til 1994 steg omsætningen fra 60 millioner dollars til 3,5 milliarder. Det var især aerobic-danseskoene, der fik væksten til at eksplodere i 1980'erne.

Reebok blev grundlagt for 100 år siden i det nordlige England af Joseph Foster. Hans sønner og sønnesønner fortsatte produktionen af sportssko i høj kvalitet og skiftede læderet ud med nye materialer og ny produktionsteknologi.

I 1979 blev amerikaneren Paul Fireman partner i virksomheden, som han helt overtog i 1982 sammen med UK Investment Group, Pentland Industries.

Fosterfamilien var repræsenteret i bestyrelsen indtil slutningen af 1980'erne, hvor Reebok blev børsnoteret i USA.

Børsen, 26. maj 1995

Højere Handelseksamen
Handelsskolernes enkeltfagsprøve
maj 2004

04-0-2

Engelsk Niveau A

Dette opgavesæt består af 3 opgaver, der indgår i den samlede bedømmelse med følgende omtrentlige vægte:

Opgave I	25%
Opgave II	25%
Opgave III	50%
I alt	100%

Grundtekst løst ilagt.

Uddannelsesstyrelsen

Torsdag den 13. maj 2004
kl. 8.30-13.30

I

Giv et resumé på dansk af den engelske tekst "Sewing a seam of worker democracy in China" (ca. 200 ord).

Der lægges vægt på en sammenhængende fremstilling med en objektiv gengivelse af tekstens hovedpunkter.

II

Oversæt nedenstående tekst til engelsk. Tekststykket kan findes indrammet i den danske tekst "Reebok fører det etiske korstog".

Der lægges vægt på en præcis oversættelse til et korrekt og flydende engelsk.

Reebok-prisen

Samme år stiftede Reebok en pris, som hvert år uddeles til et ungt menneske, der har gjort en særlig indsats for menneskerettighederne. Prisen på ca. 25.000 dollars er bl.a. givet til en af de kinesiske ledere af studenteroprøret på den Himmelske Freds Plads. I 1992 gik prisen til den dengang 12-årige pakistanske dreng, Iqbal Masih, der som fire-årig blev solgt som slavearbejder til en tæppefabrik. I seks år arbejdede han 12-14 timer i døgnet. Da han slap ud af slaveriet, rejste han rundt i forskellige lande og fortalte sin historie, hvilket har skabt debat og boykot af flere lande, som benytter slave- og børnearbejde. I februar i år blev Iqbal Masih myrdet, 15 år gammel.

Paul Fireman er formand for Reeboks fond for menneskerettigheder. [...]

Den Himmelske Freds Plads: Tiananmen Square

III

Løs nedenstående opgave A eller B. Kun én af opgaverne skal løses. I besvarelsen inddrages oplysninger fra det engelske og det danske tekstmateriale.

Der lægges vægt på en selvstændig og fyldig besvarelse.

A

Amnesty Internationals erhvervsafdeling har indbudt en række erhvervsledere til en konference om vestlige virksomheders menneskerettighedspolitik, især m.h.t. faglig organisering i tredieverdenslande.

Doug Cahn, der er ansvarlig for Reeboks menneskerettighedspolitik, er inviteret til at holde en tale om Reeboks bestræbelser på at overholde menneskerettighederne i lavtlønsområder i Asien. I talen påpeger han vigtigheden af at give medarbejderne mulighed for at organisere sig i en fagforening og begrunder hvorfor. Han giver gode råd om, hvordan sådanne valg bedst gennemføres og kommer med forslag til, hvordan de nyvalgte fagforeningsrepræsentanter kan styrkes i deres nye rolle. Han slutter af med kort at fortælle om, hvad Reebok gør for at sikre arbejdernes rettigheder.

Skriv Doug Cahns manuskript til talen på engelsk.

B

Skriv et essay om etik i forbindelse med produktion af varer og tjenesteydelser i tredieverdenslande.

Skriv essayet på engelsk, og giv det en passende overskrift.

Appendix B: Extracts from the Scoring Rubric for HHX, Written EFL Exam (scale + translation)

(originals follow the translations)

B1: Marking scale in the Danish Education System (until 1 year ago)

The following marking scale is used throughout the Danish Education System (official translation):

13: is given for the exceptionally independent and excellent performance.
11: is given for the independent and excellent performance.
10: is given for the excellent but not particularly independent performance.
9: is given for the good performance, a little above average.
8: is given for the average performance.
7: is given for the mediocre performance, slightly below average.
6: is given for the just acceptable performance.
5: is given for the hesitant and not satisfactory performance.
03: is given for the very hesitant, very insufficient and unsatisfactory performance.
00: is given for the completely unacceptable performance.

Danish version:
13: Gives for den usaedvanlig sevstaendige og udmarkede praestation
11: Gives for den udmaerkede og selvstaendige praestation
10: Gives for den udmaerkede, men noget rutinepraegede praestation
9: Gives for den gode praestation, der ligger lidt over middel
8: Gives for den middelgode praestation
7: Gives for den jaevne praestation, der ligger lidt under middel
6: Gives for den noget usikre, men nogenlunde tilfredsstillende praestation
5: Gives for den usikre og ikke tilfredsstillende praestation
03: Gives for den meget usikre, meget mangelfulde og utilfredsstillende praestation
00: Gives for den helt uantagelige praestation

(a score of 6 or above indicates that the student has passed the test in question; a score of 5 or below indicates that the student has not passed the test in question)

It is a criterion-referenced rating system.

B2: Converted scale

For the current study the scale was converted into a more manageable form:

B3: Scale conversion

Original scale	Converted scale
13	12
11	11
10	10
9	9
8	8
7	7
6	6
5	5
03	4
00	3

As no scores of 13 or 0 were given in this study (and these scores are, in fact rare), only the original score of 03 was converted to a 4.

B4: Assessment criteria for EFL exam, Part III (composition)

Danish version:
Fylde
Strukturering, Indledning, Afrunding
Anvendelse af tekstmaterialet og forstaaelse of problemstillingen)
Formuleringsevne: Ordforraad, syntaks, variation, idiomatik, grammatik
Formalia

B5: Scale descriptors for EFL exam, Part III (composition)

(my translation)
It is stressed that the scale descriptors below are to be used as a help and not to be followed mechanically and thus that the assessment is to be holistic.

Essay
Excellent (13, 11, or 10)
Fully elaborated task fulfilment, which includes not only all the points required by the task but also independent contributions
Structure of the text is clear and appropriate
Relevant introduction and conclusion
Good argumentation
If required by the task the materials available are used independently
Solid command of basic grammar
Vocabulary is advanced, varied and precise

Appropriate use of terms, words and phrases from the materials available
Fluent language. Few or no language mistakes. No mistakes leading to confusion or misunderstanding
Varied syntax and use of idiomatic expressions

Average (a score of 9, 8, or 7)
Somewhat elaborated task fulfilment, which covers all the points required by the task
Structure of the text is relatively clear and appropriate
Short and not particularly independent introduction and conclusion
Documents a fair understanding of the communication situation
Satisfactory argumentation
Materials available used to some extent but not in a structured way
Appropriate task style
Adequate command of basic grammar
Adequately advanced, varied and precise vocabulary
Language fluent to some extent.
Some grammar mistakes but few leading to confusion or misunderstanding
Syntax somewhat varied. Little (or no) use of idiomatic expressions

Passed (a score of 6)
Minimal task fulfilment, which covers most of the points required by the task
No clear structure of the text. Often no introduction or conclusion
Documents lack of understanding of the communication situation
Arguments and suggested solutions do occur, but they are not particularly elaborated or documented
Materials available used sporadically
Lack of task appropriate style
Some command of basic grammar
Vocabulary not varied, advanced or precise
Language full of mistakes but comprehensible. Some mistakes, however, lead to confusion or misunderstanding
Primarily simple syntax

Not passed (a score of 5, 03, or 00)
Task fulfilment very minimal, which does not cover all of the points required by the task
Text reveals lack of understanding of the communication situation
Irrelevant contributions to the content
Materials available have been misunderstood or are not used
Text lacks structure and coherence
No (or superficial) argumentation or documentation
Task style not appropriate
Insecure grammar
Restricted vocabulary with repetitions
Confusing language, which is full of mistakes and difficult or impossible to understand
A lot of negative transfer from Danish
Very simple syntax

Danish version:

Udmaerket (13, 11, 10)
Fyldig besvarelse, der daekker alle punkter i oplaegget samt indeholder selvstaendige bidrag.
Elevens tekst er klart og hensigtsmaessigt struktureret.
Der er en relevant indlednign og en konklusion med perspektivering til slut.
Besvarelsen er velargumenteret og inddrager viden fra sprogomraadet, hvor det er hensigtsmaessigt.
Hvor det kraeves i opgaveformuleringen, anvendes tekstmaterialet selvstaendigt og med overblik.
Solid beherskelse af den elementaere grammatik.
Varieret, nuanceret og fagligt ordforraad.
Hensigtsmaessig anvendelse af ord og vendinger i det engelske materiale.
Flydende sprog. Kun faa eller ingen sprglige fejl, og ingen af dem er meningsforstyrrende.
Varieret syntaks og idiomatiske udtryk anvendes.

Middel (9, 8, 7)
Nogenlunde beherskelse af den elementaere grammatik.
Nogenlunde varieret, nuanceret og fagligt ordforraad.
Nogenlunde flydende sprog med en del sporoglige fejl, men kun faa af dem meningsforstyrrende.
Nogenlunde vareret syntaks men kun faa eller ingen idiomatiske udtryk.

Bestaaet (6)
Nogenlunde beherskelse af den elementaere grammatik.
Ordforraadet mangler varaition, nuancering og faglighed.
Sproget er knudret og fejlfyldt, men overvejende forstaaeligt, selvom der forekommer en del meningsforstyrrende fejl. Der anvendes hovedsagelig enkle syntaktiske strukturer, der behandles rimeligt sikkert.

Ikke bestaaet (5, 03, 00)
Usikkerhed i elementaer grammatik.
Begraenset ordforraad med gentagelser.
Rodet og fejlfyldt sprog,d er er vanskeligt eller umuligt at forstaa i stoerre eller mindre dele af elevens tekst.
Mange danismer og en staerkt forenklet syntaks forekomer.

www.us.uvm.dk/gymnasie/erhverv/eksamen/Vejledninger/vejl99en.html

Appendix C: Profile Questionnaire and Summary of Profiles

Spørgeskema i forbindelse med ph.d.-undersøgelse om bedømmelsesstrategier

Personlig profil:

1. **Navn (eller pseudonym):**

   _____

2. **Dato (for udfyldning af skema):**

   _____

3. **Køn (sæt kryds): Kvinde _____ mand _____**

4. **Alder <30___ 31-40 ___ 41-50 ___ 51-60 ___ >60**

5. **Uddannelse:**

   _____

6. **Års undervisningserfaring:**

   _____

7. **Års censorerfaring:**

   _____

Rater Profile Summary

Rater name	Gender	Age	Education	Teaching experience	CWA Rating experience
Lone	Female	41-50	Cand.Mag	12	12
Louise	Female	31-40	Cand.Ling.Merc.	10	10
Thea	Female	31-40	Cand.Mag	7	3
Malene	Female	41-50	EA	12	6
Nina	Female	31-40	Cand.Ling.Merc.	7	4
Jens	Male	41-50	Cand.Ling.Merc.	25	20
Susanne	Female	51-60	Cand.Ling.Merc.	30	30
Tove	Female	51-60	Cand.Mag.+ED	21	21
Jesper	Male	41-50	Cand.Ling.Merc.	14	8
Pernille	Female	51-60	Cand.Ling.Merc.	33	5
Helle	Female	>60	Cand.Paed.	30	15
Astrid	Female	>60	Cand.Mag.	15	12
Tina	Female	51-60	Cand.Ling.Merc.	33	31
Torben	Male	41-50	Cand.Interpret.	13	13
Gitte	Female	>60	Ed + SIF	30	30
Julie	Female	51-60	Cand.Mag.	30	25
Hans	Male	51-60	Cand.Mag.	24	17
Henrik	Male	>60	Cand.Mag.	35	30
Jette	Female	51-60	Cand.Ling.Merc.	25	6
Ken	Male	41-50	Cand.Ling.Merc	12	8

Appendix D

Warm-up Exercises to TA

1. Taenk hoejt mens du forsoeger at loese foelgende regneopgave:

"Hvad er 24 gange 34"

2. Taenk hoejt mens du finder frem til hvor mange vinduer der er i dit hjem.

Translation:

1. *Think aloud while you calculate the following math problem:*

   "Multiply 24 times 34"

2. *Think aloud in your attempt to find out how many windows are in your house.*

Appendix E:

Transcription Conventions

Symbol	Explanation for symbol
*	Incomprehensible item
**	Two or more incomprehensible items
...	A pause of five seconds or more
CAPITAL LETTERS	Stretches of text read aloud.
CORRECT	Stretch of text repeated in corrected form
"from dictionary"	" " Read aloud from other texts than the scripts
' own notes'	' ' Rater quotes from own notes
mmm	Hesitation sound
uhuh	Uptaking sound
haha	laughing
ehrm	Floor keeping or floor taking gambit

Appendix F:
Samples of coded data (illustrative exemplar samples from protocols)
with translations into English

Interpretation Strategies

Contextual or Monitoring Focus

ICM1 Read or Interpret Task Input/Source Material

Og det staar SKRIV ET SCRIPT ETIK I FORBINDELSE MED PRODUKTION AF VARER OG TJENESTEYDELSER I TREDIEVERDENSLAND
Hans, script 3

Translation:
And it says WRITE AN SCRIPT ON ETHICS IN RELATION TO PRODUCTION OF GOODS AND SERVICES IN THE THIRD WORLD
Hans, script 3

ICM2 Read or Reread Composition

AN EXAMPLE TO THIS COULD BE THAT THEY WROTE FACTS ON THE SHOEBOXES ABOUT THE CONDITINS AT THEIR FACTORIES. I THINK THAT IF THE CONSUMERS KNOW ABOUT THE DEMOCRACY IN CHINA AND THEY SHOULD CHOOSE BETWEEN REEBOK AND ANOTHER SHOE COMPANY WITH SAME PRICE AND QUALITY og saa mangler der et lille komma
Henrik, script script 5

Translation:
AN EXAMPLE TO THIS COULD BE THAT THEY WROTE FACTS ON THE SHOEBOXES ABOUT THE CONDITINS AT THEIR FACTORIES. I THINK THAT IF THE CONSUMERS KNOW ABOUT THE DEMOCRACY IN CHINA AND THEY SHOULD CHOOSE BETWEEN REEBOK AND ANOTHER SHOE COMPANY WITH SAME PRICE AND QUALITY
And then a comma missing
Henrik, script script 5

ICM3 Envision Personal Situation of the Student

Ja, jeg sidder lige og kigger, taenker paa, her naar der bliver skrevet om Nike i de sidste 8 linier af andet afsnit, at det lyder meget som noget der har vaeret brugt foer. Man kunne godt forestille sig at eleven har skrevet en opgave om det her foer og sidder og genbruger fra tidligere opgaver
Malene, script 7

Translation:
Right now I am thinking that here when they write about Nike in the last 8 lines of the second paragraph, that it sounds as if it has been used before. Perhaps the student has written an assignment about this topic before and then copies from these older assignments.

Malene, script 7

ICM4 Consider Task or Exam Requirements

ha, ha. men den er jo ogsaa ehrm utrolig kortfattet opgaveformulering i forhold til hvad de plejer. Der er jo ikke nogen form for hjaelp egentlig
Malene to Thea, script 1

Translation:
Ha, ha, but it is also incredibly short the script question compared to what it usually is. It is no help for the student really.
Malene to Thea, script 1

ICM5 Consider Own Perception of Correct English (e.g. Consult a Dictionary)

Ja, saa staar der igen ETHICS HAS. Jeg tror lige for en sikkerheds skyld, at jeg vil slaa op om det ikke er rigtigt at ETHICS skal vaere i forbindelse med flertal, saa jeg ikke retter noget der er, som alligevel er rigtigt. Jeg slaar op i en engelsk-engelsk ordbog McMillan og skal finde ETHICS, og det har vi her "ethics". Der er flertal "principles that are used to decide" ehrm "what is right and what is wrong".
Ja, saa det maa skulle forbindes med flertal. Jeg laegger ordbogen vaek.
Malene, script 5

Translation:
Yes, again it says ETHICS HAS. I think I will look it up to make sure that ETHICS must be combined with a plural [form of the verb], so that I am not correcting something that is, in fact, correct. I am looking it up in a monolingual dictionary. McMillan, finding ETHICS, here we have it "ethics". It is plural "prinicples that are used to decide" ehrm "what is right and what is wrong". Yes, it must be combined with a plural [verb]. I am putting the dictionary away.
Malene, script 5

Textual Focus – Quantity

ITV1 Scan Composition for Length

Nej, lad os nu se, jeg kigger lige. Hvor lang er den, halvanden side.
Nina, script 5

Translation
No, let's see. I am looking. How long is it, a page and a half.
Nina, script 5

Textual Focus – Structure

ITSTR1 Scan Composition for Structure

IT IS VERY IMPORTANT TO ALWAYS KEEP THE HUMAN RIGHTS IN MIND. Ok, ja, der kommer en indledning her CONCERNING ETHICS
Louise, script 2

Translation
IT IS VERY IMPORTANT TO ALWAYS KEEP HUMAN RIGHTS IN MIND. OK, yes, there is an introduction here CONCERNING ETHICS
Louise, script 2

saa gaar vi over til konklusionsafsnittet
Jesper, script 6

Textual Focus – Content and Use of Source Materials

ITC1 Discern or Summarize Ideas

Saa han siger, at hvis de gerne vil vaere paa linie med human rights, hvad kan de saa goere for at forhindre deres udenlandske partnere i at fortsaette udnyttelsen.
Helle, script 12

Translation:
So he says that they want to comply with the human rights, what can they do then to prevent their third world partners from continuing the exploitation.
Helle, script 12

ITC2 Identify or Interpret Ambiguous Phrases

Hm, jeg er ikke helt klar over hvad han mener her; maaske noget med ehrm at ikke altid er klar over om der kommer nogen og tjekker dem med hensyn til hvorvidt de overholder etiske regler. Men jeg ved ikke rigtig. Jeg proever at laese videre.
Helle, script 12

Translation:
Hmm, I'm not sure what he means here; perhaps something about ehrm that not always certain that people come by to check whether they obey the ethic rules. But I don't know. I'll try to read on.
Helle, script 12

Textual Focus – Language

ITL1 Classify Language Errors into Types

Der staar WHO IS, det er en kongruensfejl
Malene, script 1

Translation:
It says WHO IS, it is a subject-verb agreement error.
Malene, script 1

ITL2 Identify Errors

REEBOK, WHO WE GOT AN ARTICLE FROM. Ja, det er jo ikke rigtigt.
Ken, script 15

Translation:
REEBOK, WHO WE GOT AN ARTICLE FROM. Well yes, this is not correct.
Ken, script 15

ITL3 Edit Language (error or unclear phrases)

THE EXTEND er stavet med 'd'. Jeg streger lige ud og retter til 't'.
Louise, script 2

Translation:
THE EXTEND spelled with a 'd'. I am crossing that out and substituting it for a 't'.
Louise, script 2

Textual Focus – Style

ITSTY1 Discern Style, Register, or Genre

Der er lidt foelelser, og det mener jeg ogsaa kommer til sidst et sted
Hans to Henrik, script 6

Translation:
There are some feelings, and I believe we also have that at the end somewhere.
Hans to Henrik, script 6

Judgment Strategies

Contextual or Monitoring Focus

JCM1 Articulate Score

Jeg har nok. Jeg har nok. Jeg vil godt vaere med paa en 7'er, for det tror jeg er rigtig nok.
Hans to Henrik, script 5

Translation:
I probably have. I probably have. I'll go along with a 7, because I think is appropriate here.
Hans to Henrik, script 5

JCM2 Compare Scripts

og ehrm, ja uden at jeg egentlig kan saette fingre paa det, saa synes jeg ikke den er saa god som 4'eren.
Jesper, script 5

Translation:
And ehrm, yes without knowing exactly why I don't think it is as good as script number 4.
Jesper, script 5

JCM3 Define, Revise, or Suggest Assessment Strategies

ja, og der har jeg nok, hvis vi saa skal tage en samlet konklusion, der har jeg nok fokuseret lidt for meget paa det der indhold. Jeg blev simpelthen irriteret
Gitte to Julie, script 3

Translation:
Yes, I probably have, so if we have to reach a conclusion, I probably have focused too much on the content. I just so annoyed.
Gitte to Julie, script 3

naa, ja, der har jeg godt nok ikke vaeret nede der. det er nok en af mine fejl. det er at jeg simpelthen saa bange for at give de her karakterer.
Pernille to Jesper, script 9

Translation:
Well, yes I didn't go that far down. It is probably one of my mistakes. I'm simply too afraid to assign these scores.
Pernille to Jesper, script 9

Jeg kan godt. Jeg synes retoriske spoergsmaal er ok. Men jeg synes at man er noedt til at redegoere inden man kan begynde at stille de der
Torben to Tina, script 3

Translation:
I can do that. I think rhetorical questions are ok. But I think that they must elaborate some ideas before they ask them
Torben to Tina, script 3

JCM4 Articulate General Impression

saa alt i alt er det en paen og ordentlig opgave
Hans, script 12

Translation:
Soo all in all a nice and neat paper
Hans, script 12

JCM5 Deliberate/Articulate Teaching Practices

og der ville jeg saa sige til mine egne elever at at det det er helt fint hvis de kan supplere med noget andet
Malene to Thea, script 15

Translation:
In this situation I would say to my students that it is absolutely fine if they can supply with other information.
Malene to Thea, script 15

JCM6 Exemplify Directly from Student Text

meget mere, og saa har jeg sat rosende stjerner ehrm over for, over for forskellige ting og sager, for eksempel, der er noget her THE CANDIDATES i tredie afsnit; elevens tredie. THE CANDIDATES PREPARE SPEECHES TO CONVINCE THE WORKERS IN THE BATTLE TO WIN AND IMPROVE. Det synes jeg har en retorisk vaerdi i sig, ikke?
Astrid to Helle, script 2

Translation:
Much more, and then I have praised it with stars here and there, for instance something here THE CANDIDATES in the third paragraph; the student's third. THE CANDIDATES PREPARE SPEECHES TO CONVINCE THE WORKERS IN THE BATTLE TO WIN AND IMPROVE. I think it has a lot of rhetorical value, doesn't it?
Astrid to Helle, script 2

og ten og TEEN, blander hun ogsaa sammen, ikke?
Astrid to Helle, script 9

Translation:
And she also mixes up ten and TEEN, doesn't she?
Astrid to Helle, script 9

JCM7 Consider official or consensus-based Strategy

og saa kan man sige at vi plejer jo at goere det at vi ehrm saa lader det komme eleven til gode, saa vi giver den hoeje karakter.
Malene to Thea, script 1

Translation
And then we could say what we, in fact, usually do ehrm we usually give the student the benefit of the doubt, so we assign the higher score.
Malene to Thea, script 1

JCM8 Consider Personal Response or Bias

men jeg bliver bare enormt irriteret naar de, ha, ha, laver saadan noget.
Malene to Thea, script 10

Translation:
But I just get so annoyed when they ha, ha, when they do things like that.
Malene to Thea, script 10

Textual Focus – Quantity

JTV1 Assess or Justify Quantity

saa derfor synes jeg at det, men den er saa ikke meget fyldig, kan man sige, og derfor saa kan den aldrig komme hoejere op.
Malene to Thea, script 13

Translation:
So that's why I why, but you could say that it is not so long, and that's why the score isn't higher.
Malene to Thea, script 13

Textual Focus – Structure

JTSTR1 Assess Structure Overall

... Jeg har faktisk skrevet at den her er ikke ordentligt struktureret i den sidste halvdel. Der er bre saadan en lang smoere med uklare afsnitsinddelinger.
Jens to Nina, script 4

Translation:
I have, in fact, written that this one does not have a good structure in the last part of the script. There is bre this long passage of unclear divisions of the paragraphs.
Jens to Nina, script 4

JTSTR2 Assess or Justify Headline

CHEAP LABOUR – THE FLIP SIDE OF GLOBALISATION? Overskriften faar et '+' i margen. rammende overskrift. Den siger noget om hvad det hele handler om
Louise, script 4

Translation:
CHEAP LABOUR – THE FLIP SIDE OF GLOBALISATION? The title gets a '+' in the margin. Catching title. It tells us what it is all about.
Louise, script 4

JTSTR3 Assess or Justify Introduction and/or Conclusion

Man skulle skrive et essay om, i forbindelse med produktion af varer og tjenesteydelser i tredieverdenslande. Saa der er da en en pudsig ehrm indledning. Hvis man da overhovedet kan kalde det en indledning
Malene, script 14

Translation:
They are supposed to write an essay on the production of goods and services in the third world. So this is an odd introduction. If you can call it an introduction at all.
Malene, script 14

JTSTR4 Assess or Justify Coherence and/or Cohesion

Der kunne ogsaa maaske have vaeret en bedre overgang mellem foerste og andet afsnit ehrm.
Jesper, script 3

Translation:
Also, there could have been a better transition from the first paragraph to the second one ehrm.
Jesper, script 3

Textual Focus – Content and Use of Source Materials

JTC1 Assess Content/Ideas Overall

Det var en ok betragtning. Jeg laver et '+' i margen og skriver der '+ indh.' ude i margen
Louise, script 7

Translation:
That was an okay perspective. I am putting a '+' in the margin and I am writing '+ introduction' in the margin.
Louise, script 7

JTC2 Assess or Justify Reasoning, Logic, or Topic Development

Det er en god taktisk ide, synes jeg, at stille spoergsmaal, hvis man gerne vil have ordet selv. * kan man faa svaret, og det er det eleven laegger op til. Men eleven kommer allerede i tredie linie med et svar, som jeg synes er en lille smule overfladisk, fordi de ikke vil have daarlig samvittighed. Jeg ved ikke helt om om man allerede saa hurtigt kan besvare det. Her saetter jeg en lille boelgestreg ehrm. Jeg er ikke ehrm tilfreds med argumentet.
Astrid, script 3

Translation:
*It is a good idea, tactically, I think, to ask questions if you want the word * you can get the answer, and that's what the student is trying to do. But the student comes up with an answer*

already in the third line, which I think is a bit superficial, because they don't want a bad conscience. I don't really know if one could answer this so soon. I will underline this ehrm. I am not satisfied with the argument.
Astrid, script 3

WE CANNOT, BECAUSE WHY SHOULD THE COMPANIES WHICH DO NOTHING TO IMPROVE THE CONDITIONS, SHOW IT IN PUBLIC? Men der jo andre end selskaber der toer oplyse om det her. Saa det er ikke helt logisk argumenteret.
Astrid, script 5

Translation:
WE CANNOT, BECAUSE WHY SHOULD THE COMPANIES WHICH DO NOTHING TO IMPROVE THE CONDITIONS, SHOW IT IN PUBLIC? But other companies do, in fact, come forth with this. So it is not a logic argument.
Astric, script 5

JTC3 Assess or Justify Clarity
Ja, og daarlig forstaaelse, saa tror jeg ogsaa at der er tvivl om hvad der er meningen egentlig var med den saetning der, hvad det var man proevede"
Thea to Malene, script 15

JTC4 Assess or Justify Correctness of or Disagreement with Content
Hvad for noget? IT WILL BE HARD TO HAVE BIG FACTORIES WITHOUT THE ETHIC RULES det er det da ikke. Den kan man da sagtens. THE ETHICS RULES. Streg under.
Thea, script 11

Translation:
What? IT WILL BE HARD TO HAVE BIG FACTORIES WITHOUT THE ETHIC RULES not it is not. They can easily do that. THE ETHICS RULES. Underline.
Thea, script 11

WHICH AT LEAST MUST MATCH THE LEVEL, som, naa jo. Det har du ret i. Den mindst skal matche den som den lokale fagforening ehrm ja saa saetter.
Helle, script 2

Translation:
WHICH AT LEAST MUST MATCH THE LEVEL, which, well, yes. You are right. At least it must match the one set by the local union.
Helle, script 2

JTC5 Assess or Justify Maturity or Independence
Ja, I HOPE THIS PROBLEM WILL BE HANDLED WITHIN A FEW YEARS, SO THESE CHILDREN CAN HAVE A NORMAL LIFE. Der vil jeg sige, det er naivt det her. Ehrm og naivt og og unuanceret.
Malene, script 8

Translation

Yes, I HOPE THIS PROBLEM WILL BE HANDLED WITHIN A FEW YEARS, SO THESE CHILDREN CAN HAVE A NORMAL LIFE. Here I would say that it is naïve. Ehrm and naïve and and unnuanced.
Malene, script 8

den er lidt banal. Ehrm og der er ikke den store modenhed i den, men altsaa
Gitte to Julie, script 8

Translation:
It is a bit banal. Ehrm and not much maturity in it, but well
Gitte to Julie, script 8

JTC6 Assess or Justify Task/Topic Relevance or Completion
Jeg synes simpelthen at at den er dumpet fordi at man ehrm besvarer ikke den opgave der er stillet
Malene to Thea, script 15

Translation:
I just think it cannot pass because he doesn't answer the essay question
Malene to Thea, script 15

Det kommer til at koere med ud med meget 9/11 og hvad det egentlig har betydet, som egentlig ikke har nogen relevans til det her emne
Susanne, script 3

Translation:
It goes a lot in the direction of 9/11 and the impact of that, which really has no relevance to this topic
Susanne, script 3

JTC7 Assess or Justify Use and Understanding of Source Materials
DESTROYED, AND THE WOULDN'T USE THE COMPANY ANY MORE. Meget primitiv fremstilling af det der staar i teksten.
Astrid, script 9

Translation:
DESTROYED, ANDTHE WOULDN'T USE THE COMPANY ANY MORE.
Very primitive account of the text
Astrid, script 9

Ja, der bliver i hvert fald brugt en del fra den engelske artikel her, og det er saa meget fint.
Louise, script 2

Translation:
Yes, they do use the English text a lot here, and this is very nice.
Louise, script 2

Textual Focus – Language

JTL1 Assess Language Overall

Ja, jeg sidder og ser og taenker paa at den her opgave har ogsaa et rigtigt fint sprog
Malene, script 6

Translation:
Yes, I am thinking that this essay has a very nice language
Malene, script 6

JTL2 Assess or Justify Frequency of Errors

Sprogligt har den meget faa fejl
Gitte to Julie, script 4

Translation:
Looking at the language, there are few errors
Gitte to Julie, script 4

JTL3 Assess or Justify Gravity of Errors

IT WOULDN'T SEEMS; den er jo godt nok grov den der fejl
Pernille, script 5

Translation:
IT WOULDN'T SEEMS; now that is a severe error.
Pernile, script 5

JTL4 Assess or Justify Syntax or Morphology

ehrm, der er jo flere paene konstruktioner. Den der DUE TO THE LACK OF WORKING POSSIBILITIES. Det er jo, det er jo fornuftigt nok sprogligt.
Hans, script 2

Translation:
Ehrm, well there are several nice constructions. This DUE TO THE LACK OF WORKING POSSIBILITIES. Well, this is, this is a nice language.
Hans, script 2

Ok. Det virker som om der er styr paa EASILY, TEMPORARY, osv. Noget med adjektiver og adverbier, laver et '+' i margen.
Louise, script 2

Translation:
Okay. It seems as if they command EASILY, TEMPORARY, etc. Something about adjectives and adverbs, putting a '+' in the margin.
Louise, script 2

JTL5 Assess or Justify Lexis

paent sprog for OUTSOURCED; det er et fint ord at bruge, et relativt nyt begreb, og han bruger det korrekt her. Det tyder paa at hun har antennerne slaaet ud. Det er udmaerket.

Astrid, script 12

Translation:
Nice language for OUTSOURCED; it is a nice word to use, a relatively new term, and he is using it correctly here. It is a sign that he is perceptive. It is excellent.
Astrid, script 12

JTL6 Assess or Justify Fluency or Comprehensibility
THAT HAS PUT GREAT EMPHASIS ON GIVING THE WORKERS THE BEST CONDITIONS. THEY MAKE SURE THAT THE WORKERS ARE TREATED IN THE SAME WAY AS THE ONES WHO WORK AT THE FACTORIES BASED IN AMERICA Meget, meget fin fluency i det der
Hans, script 2

Translation:
THAT HAS PUT GREAT EMPHASIS ON GIVING THE WORKERS THE BEST CONDITIONS. THEY MAKE SURE THAT THE WORKERS ARE TREATED IN THE SAME WAY AS THE ONES WHO WORK AT THE FACTORIES BASED IN AMERICA Very, very nice fluency in this one.
Hans, script 2

Altsaa man kan jo sige generelt om sproget her at der er jo saadan meget klodset. Det er lidt svaert at forstaa ind i mellem
Jette, script 14

Translation:
Well, you could say that generally the language is, in fact, very clumpsy. It is at times difficult to understand.
Jette, script 14

JTL7 Assess or Justify Spelling
DEFINITELY; som er stavet korrekt; naesten kryds for det ogsaa
Astrid, script 6

Translation:
DEFINITELY; which is spelled correctly; almost check for that one too.
Astrid, script 6

JTL8 Assess or Justify Punctuation
TO BE ABLE TO SUPPORT THE FAMILY; forvirrende tegnsaetning hele vejen igennem.
Astrid, script 8

Translation:
TO BE ABLE TO SUPPORT THE FAMILY; confusing punctuation all the way through
Astrid, script 8

Textual Focus – Style

JTSTL1 Assess or Justify Style

WE, jeg kan ikke, jeg bryder mig heller ikke om en WE-form i et essay, som er en personlig kommentar
Nina to Jens, script 1

Translation:
WE, I cannot, I don't like the WE-form either in an essay, which is a personal comment.
Nina to Jens, script 1

Men altsaa, jeg kan vaeldig godt lide hendes brug af retoriske spoergsmaal. Altsaa paa den maade. Det er nogle ting ikke ogsaa, som jeg synes.
Tina to Torben, script 3

Translation:
But I really like her use of rhetorical questions. You know, in a way. There are some things, right that I think.
Tina to Torben, script 3

JTSTL2 Assess or Justify Genre

men jeg synes ogsaa netop pointen at det ikke helt har karakter af essay.
Hans to Henrik, script 8

Translation:
But I also think that the point is that it is not the essay genre.
Hans to Henrik, script 8

Appendix G:
Retrospective Questionnaire (perceptions of score accuracy and CWA in general)

1. Hvis du ser tilbage på de stile du netop har bedømt, hvilke karakterer er du så mest tilfreds med, dine egne karakterer fra da du bedømte stilene individuelt eller de karakterer der fremkom af de fælles bedømmelser? Vurder for hver stil (sæt kryds).

Stil nr.	Egen bedømmelse	Fælles bedømmelse	Gav samme karakter
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			

2. Hvilke fordele og ulemper mener du der er i forbindelse med fællesbedømmelser?

Translation:

1. *Looking back at the essays you have rated, which scores do you find more accurate: your own independent scores or the communal scores? Assess scores for each essay.*

2. *What in your opinion are the advantages or disadvantages of communal writing assessment procedures?*

Appendix H: Full Range of Scores in Independent Ratings and in Communal Rati

Scores in Independent Ratings

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Pernille	6	8	7	8.7	8	10	9	7	6	5	5	8.3	8.5	5.7
Jesper	6	5.5	9.7	8.3	8	9	10	5	4	4	4	10	9.3	4.5
Gitte	7	8	8	9	7	10	9	7.3	5	4.3	5	9	9.3	6
Julie	6.3	8.5	9.3	9.7	7	8.7	9.7	7.7	5.7	4	4	8.3	9	5.7
Torben	6.7	8.7	7.5	9.3	8	9	8	6	5	4.7	5	8	8	5
Tina	5.5	8.5	8.3	9.5	8.8	9	9.7	8	4.7	6.5	4.7	8.7	8	5
Tove	5.5	7.3	6.7	9	8.5	8	8.5	5.3	4.5	4.7	4.5	7.3	8.5	5
Susanne	5.3	7.3	7	9.3	8.3	8.7	9.7	5.3	4.7	5.3	4.5	6.7	8.5	5.3
Nina	6.5	8	7	8	8	8	8.5	6	5.7	5	5	8.5	8.3	5
Jens	6.7	8.5	8	9	7	8.5	8.3	5.7	4.7	4	5	7.5	8.5	4.5
Lone	6	8	8	8.5	8	8.5	9	7	5	4	4	7.5	9	4
Louise	7	8.5	9.5	10	9	9	8.5	7	5.7	4.5	4.3	9	9.3	5.7
Astrid	6	8	8	10	7	10	8.5	5	4	4	4	7	8.5	4
Helle	6	7.7	7	8	8.7	9	9	6	5	4.3	5	8	8	5
Thea	7	8.3	6.7	9.3	8	10.3	9	8	5	5.7	5	7.3	8	5
Malene	6	8.3	7.7	9.3	10	9	10.3	7	5	4.5	5	6	8	5.5
Jette	7	8	8.3	9	8	9.7	10	7	5.7	5	5.5	9	8.7	5.5
Ken	6.3	8	9	9.7	8.3	8.7	9.3	7	5	5	4.7	9.7	9.7	5.3
Hans	6	7.7	7	9	6	8	9	5.3	5	4	5	8	9	5
Henrik	7.5	8	8.7	9	8.5	8	9	7	5	5	5.7	9	10	6
AVERAGE	6.3	7.9	7.9	9.1	8.0	9.0	9.1	6.5	5.0	4.7	4.7	8.1	8.7	5.1
STDEV	0.6	0.7	0.9	0.6	0.9	0.7	0.6	1.0	0.5	0.7	0.5	1.0	0.6	0.6

Scores in Communal Ratings

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Pernille +J	6	7	8	9	8	9	9	5	4	4	4	9	9	4
Gitte + Jul	7	8	9	9	7	10	9	8	5	4	4	9	9	6
Torben + 1	6	9	8	9	8	9	9	7	5	5	5	9	8	5
Tove +Sus	5	7	7	9	8	9	9	5	5	5	4	7	9	5
Nina + Jen	7	8	7	9	8	8	8	6	5	4	5	8	8	5
Lone + Lo	7	9	9	9	8	9	9	7	5	4	4	8	9	5
Astrid + H	6	8	8	10	7	9	8	5	4	4	4	7	8	4
Thea + Ma	7	8	7	9	8	10	9	7	5	5	5	7	8	5
Jette + Ke	7	8	9	10	8	9	10	7	5	5	5	10	9	5
Hans + He	7	8	7	9	7	8	9	6	5	5	6	9	10	5
AVERAGE	6.5	8	7.9	9.2	7.7	9	8.9	6.3	4.8	4.5	4.6	8.3	8.7	4.9
STDEV	0.7	0.7	0.9	0.4	0.5	0.7	0.6	1.1	0.4	0.5	0.7	1.1	0.7	0.6

Note: The reason for the decimals in the independent ratings is that some raters did not give an exact score in th

For instance, they would say "a high four", "a low six", or "a nine or a ten", thus the score 5.3 is a numerical refle
the score 6.7 a reflection of "a low seven", and the score 5.5 a reflection of "a five or a six".

ngs

15
6
5
7
6.7
7
6.5
5
5.3
7
5
6
7
5
6
7.5
5
6
6
6
6

6.1
0.8

15
5
7
7
5
6
6
5
5
6
6

5.8
0.8

ıeir independent ratings.

ction of “a high five”,

Appendix I: Rater Perceptions of CWA in General (all rater comments)

(Translation below)

	Offer opportunity to reach the most accurate score possible	**Offer opportunity to refine assessment strategies in general**	**Ensure assessment by same standards**
Pernille			- Det er vigtigt, fordi censorer skel helst have så meget fælles grundlag som muligt, og der er mange ting man ikke kan opstille regler for. Man kan ikke sådan bare give et tal. Derfor er samtale vigtigt.
Jesper	- At ens egne idiosynkrasier/foretrukne ting ikke får for meget vægt.		- At få et fælles grundlag for bedømmelse
Gitte	- 2 par øjne ser mere end eet par. Det er indlysende, at karaktererne bliver mere retfærdige. Det er let at overse fejl. - En anden ting er vægtningen af sprog og indhold i forhold til hinanden. Man kan fokusere så meget på gram.fejl, at man faktisk glemmer, at eleven har forsøgt at formulere nogle komplicerede sætninger indholdsmæssigt, så der kommer ekstra mange fejl.		
Julie	- Ved retning fokuserer vi måske på forskellige ting (sprog/indhold/struktur) og den fælles evaluering vil sikre at en persons 'kæpheste' ikke bliver altafgørende.		
Torben	- En sikkerhed så åbenlyse fejlbedømmelser undgås		
Tina	- Større garanti for		

	karakteren – 4 øjne ser bedre end 2		
Tove	- Man opnår en retfærdig karakter når man ikke kun bedømmer selv		
Susanne	- Det sker da også, at man med en enkelt opgave får revideret sin bedømmelse, fordi man måske er blevet forstyrret	- Jeg synes, det er meget vigtig med fælles bedømmelser, da vi alle har brug for at prøve vores bedømmelser af med andre.	
Nina		- Formålet med fællesbedømmelser er, at begge har mulighed for at få en forståelse for, hvordan den anden bedømmer opgaver	- Sikrer ens og fair bedømmelse landet over.
Jens	- Fire øjne bedre end to	- Ens egen opdatering i forhold til andre vurderingsniveau	
Lone	- At sikre at svipsere ikke finder sted. Man checker hinanden		- At opnå en national standard, at niveauet bliver så ens som muligt
Louise	- Sikring af korrekt bedømmelse		- Sikring af en ensartet bedømmelse
Astrid		- En bekræftigelse af ens egen dømmekraft at opleve at medcensor bedømmer some en selv.	
Helle	- At få en refærdig, objektiv bedømmelse, for man kan have haft en dårlig dag		
Thea	- Elevens sikkerhed: censorerne supplerer hinanden i rettelserne. Derved mindre risiko for fejl, større garanti for at den "rigtige" karakter gives		
Malene	- Censorerne supplerer		

	hinanden og sikrer dermed størst mulighed for den 'rigtige' karakter		
Jette	- Bliver ellers for subjektivt -4 øjne ser bedre end 2		
Ken	- 4 øjne ser bedre end 2 -Alle har kæpheste som ville domineer hvis kun een censor		
Hans	- At "vende bøtten" og se opgaven igennem igen med friske øjne og få diskuteret vægtningen af de plusser og minusser, der ligger til grund for opgaverne.		
Henrik	- Ud fra forskellige udgangspunkter kommer man frem til den bedste "fælles nævner". - Alle aspekter (idiomatik, grammatik, stilistik, indhold) får mulighed for at blive inddraget		
Number of raters who commented on this	17	4	5

Translation:

	Offer opportunity to reach the most accurate score possible	**Offer opportunity to refine assessment strategies in general**	**Ensure assessment by same standards**
Pernille			- It is important because raters must judge by the same standards. There are so many things you cannot make rules for. You cannot just assign a score like that. That is why conversation is important" (Pernille).

Jesper	- That one's own idiosyncracies/preferred aspects don't carry too much weight		- To ahieve a common framework for assessment
Gitte	- 2 eyes see more than one pair. It is obvious that the scores become fairer. It is easy to overlook errors. - Another thing is the importance attached to the content and language of the scripts. You can focus so much on grammatical errors that you actually forget that the student has attempted to construct complex sentences about some idea and so makes a lot of errors.		
Julie	- When we rate, we might look at different features (language/content/structure) and the communal ratings will ensure that one rate's idiosyncracies do not count.		
Torben	- Ensures that obvious mistakes are avoided.		
Tina	- Greater warranty for the score – 4 eyes see more than two.		
Tove	- You achieve a fair score when there is more than one rater.		
Susanne	- it does happen that one has to revise one's assessment because one has overlooked things, was distracted	- I think communal ratings are important. We all need to test our assessments against others.	
Nina		- The purpose of communal ratings is that each rater has an opportunity to gain insight into the other rater's	- Ensures common and fair assessment throughout the country.

		assessment strategies.	
Jens	- Four eyes better than two	- revision of ones' own strategies with others' level of leniency.	
Lone	- To avoid mistaken judgments. We check each other.		- To obtain a national, common standard.
Louise	- Ensure a correct assessment.		- Ensure a common assessment
Astrid		- Confirming ones' on assessments when other raters have the same assessment strategies.	
Helle	- To obtain a fair and objective judgment, because one could have had a bad day.		
Thea	- Security for the student.: The raters supplement each other in their assessments. In this way fewer risks of mistakes, bigger warranty that the 'correct' score is assigned.		
Malene	- The raters supplement each other and ensure the best possibility the "correct" score.		
Jette	-The assessments will otherwise be too subjective. -4 see more than 2		
Ken	- 4 see more than 2 - Everybody has their idiosyncracies which would dominate if there had only been one rater.		
Hans	- To revisit and look at the script again with fresh eyes and to discuss the weight of the good and bad points about the scripts.		
Henrik	- Reach a common score		

	form different perspectives. – All aspects (idiomatic, grammar, content) will be assessed.		
Number of raters who commented on this	17	4	5