UMEÅ UNIVERSITY

Does language matter?

Sources of inequivalence and demand of reading ability of mathematics tasks in different languages

Frithjof Theens

Department of Science and Mathematics Education
Umeå 2019

Dissertation for PhD
ISBN: 978-91-7855-091-3
ISSN: 1650-8858
Series title: Doktorsavhandlingar i pedagogiskt arbete
Cover design inspired by Dr. Dietmar Theens
Electronic version available at: http://umu.diva-portal.org/
Printed by: CityPrint i Norr AB
Umeå, Sweden 2019

Det ordnar sig, och gör det inte det så går det bra ändå.

Stefan Holm

Table of Contents

Abstract

Practicing mathematics is not possible without the use of language. To communicate mathematical content, not only words in natural language are used but also non-verbal forms of communication such as mathematical symbols, graphs, and diagrams. All these forms of communication can be seen as part of the language used when doing mathematics. When mathematics tasks are used to assess mathematical competence, it is important to know how language can affect students' possibility to solve the task. In this thesis, two different but related aspects of the relation between language and mathematics tasks are investigated. The first aspect concerns linguistic features of written mathematics task that can make the task more difficult to read and/or to solve. These features may result in unnecessary and unwanted reading demands, that is, the task then partially assesses students' reading ability instead of their mathematical ability. The second aspect concerns differences between different language versions of mathematics tasks used in multilanguage assessments. These differences may cause inequivalence between the language versions, that is, the task may be more difficult to solve for students of one language group than students of another. Therefore, the purpose of this thesis is to investigate some of the effects that language can have on written mathematics tasks, in particular, on the validity of mathematics assessments. The thesis focuses on unnecessary reading demands and inequivalence in multilanguage assessments. The data in this thesis are obtained from tasks of the Programme for International Student Assessment (PISA) 2012. The task texts and the student results on these tasks are analyzed quantitatively to identify the occurrence and possible sources of unnecessary reading demands and inequivalence. Think-aloud-protocols and task-based interviews of students who had worked with some of the tasks, serve to qualitatively identify possible sources of reading demands and inequivalence, respectively.

The results showed both unnecessary reading demands and inequivalence in some of the tasks. Some linguistic features were identified as possible sources of these reading demands, while others were not related to them. For example, sentence *length* was not related to reading demands of tasks in Swedish, whereas sentence *structure* was identified as a possible source of unnecessary reading demands. Some linguistic differences between different language versions of mathematics tasks were also identified as possible sources of inequivalence, and in addition there were curricular differences that were such potential sources. The findings of this thesis have implications for designing mathematics tasks both in one language and in multilingual settings. They may help to ensure validity of mathematics assessments, but also to make mathematics texts easier to understand for students in general.

List of papers

I. Bergqvist, E., Theens, F., Österholm, M. (2018). The role of linguistic features when reading and solving mathematics tasks in different languages. *Journal of Mathematical Behavior*, Elsevier 2018, Vol. 51: 41-55

II. Theens, F. (2019). *Variations in Students' Reading Process when Working on Mathematics Tasks with High Demand of Reading Ability*. Paper presented at the Eleventh Congress of the European Society for Research in Mathematics Education (CERME11), Utrecht, Netherlands, February 6-10, 2019. Manuscript accepted for publication.

III. Theens, F., Bergqvist, E., Österholm, M. (preprint). T*he relation between linguistic features and DIF in multilanguage mathematics assessment*s.

IV. Theens, F. (preprint). *Using Students' Reflections to Identify Sources of Inequivalence in Translated Mathematics Tasks*.

The published papers are reproduced with permission of the relevant publisher.

Study I was conducted together with Bergqvist and Österholm. All three authors have participated in designing, conducting, and writing the study and contributed equally to the work. My main contributions were work with the analysis of the German tasks, the statistical analyses, the examples, and the background about the linguistic features.

Study III was conducted together with Bergqvist and Österholm. All three authors have participated in designing, conducting, and writing the study. I am the first author, since I had a bigger part in designing the study, conducted all statistical analyzes, and did a big part of the writing.

Abbreviations

DIF — Differential Item Functioning
DRA — (unnecessary) Demand of Reading Ability
ENG — English
GER — German
IRT — Item Response Theory
OECD — Organisation for Economic Co-operation and Development
PCA — Principal Component Analysis
PCAP — Pan-Canadian assessment program
PISA — Programme for International Student Assessment
SWE — Swedish
TAP — think-aloud protocol

Enkel sammanfattning på svenska

Språket spelar en viktig roll när man bedriver matematik. Matematiskt innehåll kommuniceras med ord i naturligt språk men även med icke-verbala former av kommunikation, såsom matematiska symboler, grafer eller diagram. Dessa kan då ses som en del av språket som används i matematiska sammanhang. Att veta hur språket som används i en matematikuppgift kan inverka på elevers möjlighet att lösa uppgiften är viktigt bland annat när uppgifter används i bedömning av matematikkunskaper.

Det övergripande syftet med denna avhandling är därför att undersöka några möjliga effekter som språket kan ha på skriftliga matematikuppgifter och därmed också på validiteten vid bedömning av matematikkunskaper. I avhandlingen behandlas två olika men besläktade aspekter formulerade som två övergripande frågor. Den första frågan handlar om vilka språkliga egenskaper i en matematikuppgifts text som påverkar svårighetsgraden att läsa och lösa uppgiften, till exempel den genomsnittliga ordlängden eller meningslängden. Denna fråga undersöks för matematikuppgifter på svenska, engelska och tyska och behandlas i artikel I och II som ingår i denna avhandling. Den andra övergripande frågan handlar om ekvivalens mellan en matematikuppgifts olika språkversioner som ges till elever på deras respektive hemspråk. Målet är att hitta möjliga orsaker för inekvivalens mellan olika språkversioner av matematikuppgifter. Det är både lingvistiska skillnader i uppgiftstexten som undersöks men också andra möjliga hot mot ekvivalensen. Denna fråga behandlas i artikel III och IV i denna avhandling. Även där undersöks och jämförs matematikuppgifter på svenska, engelska och tyska. Mer detaljerade frågeställningar angående dessa två aspekter finns i de fyra artiklarna som ingår i avhandlingen.

Första aspekten: Lingvistiska egenskaper som är relaterade till svårigheter att läsa och lösa matematikuppgifter

För att en elev ska kunna lösa en skriftlig matematikuppgift krävs det förutom matematisk förmåga alltid även en viss nivå av läsförmåga hos eleven för att kunna läsa och förstå texten och kunna lösa uppgiften. Uppgiftstexter kan variera från att innehålla endast matematiska symboler (t ex $3 + 5 = ?$) till längre texter i naturligt språk. (T ex. ”Evelina har tre marsvin och fem kaniner. Hur många djur har hon sammanlagt?”) Eftersom det ingår i matematisk kompetens att kunna kommunicera matematik, kan en del läskompetens anses tillhöra matematikkompetensen. Ord och uttryck med matematiskt innehåll (som t ex *addera*, *lutning* eller *derivata*) ingår i ett matematiskt språk som behövs för en sådan kommunikation. Andra lingvistiska egenskaper som ställer krav på läsförmåga (t ex ovanliga, icke-matematiska ord eller onödigt komplicerad meningsbyggnad) kan

dock inte anses vara en del av matematisk kompetens. De senare egenskaperna bör därför undvikas i uppgifter som är avsedda för att mäta elevernas matematiska kompetens. Annars finns risken att uppgiften till viss del också mäter läskompetensen utöver den del som ingår i matematisk kompetens.

För att undersöka sambandet mellan några lingvistiska egenskaper och matematikuppgifters onödiga krav på läsförmåga (demand of reading ability – DRA) analyserades i studie I matematikuppgifter som ingick i PISA-undersökningen 2012 på svenska, engelska och tyska. För varje språk mättes DRA med hjälp av en Principalkomponentanalys (PCA) och den allmänna svårighetsgraden bestämdes med hjälp av lösningsfrekvensen. I ett nästa steg söktes det efter möjliga korrelationer mellan DRA respektive svårighetsgrad och fyra lingvistiska egenskaper, nämligen ord-, menings- och textlängd samt informationstäthet. Resultaten visar inga statistiskt säkra korrelationer för uppgifter på engelska överhuvudtaget. Inte heller kunde det påvisas några korrelationer mellan de lingvistiska egenskaperna och DRA för uppgifter på svenska. De enda statistiskt säkra sambanden finns för uppgifterna på svenska och tyska mellan svårighetsgrad och informationstäthet samt för de tyska uppgifterna mellan DRA och både informationstäthet och ordlängd.

För att ytterligare undersöka möjliga orsaker för DRA fick svenska elever i studie II arbeta med två av de uppgifter som hade uppvisat hög DRA. Eleverna tänkte och läste högt under arbetet. På det sättet identifierades variationer i elevernas läsprocesser, såsom stamning eller upprepning. Dessa variationer pekar på möjliga orsaker för det förhöjda kravet på läsförmåga i uppgifterna, t ex ovanliga, icke-matematiska ord och komplex meningsbyggnad.

Andra aspekten: Källor av inekvivalens mellan olika språkversioner av matematikuppgifter

Det förekommer ibland att matematikuppgifter används i sammanhang där uppgifterna ges till elever på olika språk. Det kan till exempel vara de svenska nationella proven i matematik som även ges på engelska till elever som har haft engelska som undervisningsspråk. Det gäller också nationella mätningar av matematikkunskaper i flerspråkiga länder som till exempel i Kanada. Men även i multinationella kunskapstester som PISA (Programme for International Student Assessment) eller TIMSS (Trends in International Mathematics and Science Study) måste uppgifterna översättas till många olika språk. För att inte äventyra validiteten i sådana multinationella kunskapstester är det viktigt att uppgifternas olika språkversioner är så ekvivalenta som möjligt. Det innebär framför allt på funktionell nivå att de olika versionerna mäter samma innehåll med en jämförbar svårighetsgrad.

I studie III undersöktes om skillnader för tre lingvistiska egenskaper (användningen av aktiv/passiv form, användningen av andra/tredje person, olika meningskomplexitet) har samband med inekvivalens på funktionell nivå mellan olika matematikuppgifters språkversioner i svenska, engelska och tyska. För att identifiera sådan inekvivalens mellan språkversionerna genomfördes en statistisk analys av ”Differential Item Functioning” (DIF) för matematikuppgifterna från PISA 2012. Dessutom jämfördes förekomsten av de tre lingvistiska egenskaperna i språkversionerna. Resultaten visar att det förekommer DIF, dvs. vissa språkversioner är inte ekvivalenta. Mest DIF förekommer mellan den svenska och den engelska versionen. Skillnader i användningen av de tre lingvistiska egenskaper kunde också påvisas, där den vanligaste skillnaden var i användningen av aktiv och passiv form. Det finns dock inga statistiskt säkra korrelationer mellan DIF och de lingvistiska skillnaderna. En möjlig orsak till denna avsaknad av korrelationer kan vara att eleverna i den åldern som PISA-uppgifterna är gjorda för (15 år) har en så pass bra språklig förmåga att de undersökta lingvistiska egenskaperna inte påverkar deras förmåga att lösa matematikuppgifterna nämnvärt.

För att identifiera möjliga orsaker till inekvivalens mellan olika språkversioner av en matematikuppgift gjordes i studie IV en kvalitativ undersökning med hjälp av intervjuer med elever från Sverige och Tyskland som hade arbetat med några av uppgifterna som i studie III hade identifierats att uppvisa DIF mellan den svenska och tyska versionen. I denna studie identifierades, förutom skillnader i lingvistiska egenskaper (användning av ovanliga ord, formulering av vissa meningar), även andra möjliga skillnader som skulle kunna vara orsaker till inekvivalensen mellan de svenska och tyska versionerna av uppgifterna. En skillnad i hur elever från de olika språkgrupperna upplevde uppgifterna var att de tyska eleverna påverkades mera av information i texten som inte var nödvändig för att lösa uppgiften. Det gällde både förklarande text med bakgrundsinformation och kvantitativa värden som inte behövdes i beräkningen. Kulturella skillnader mellan de svenska och tyska eleverna kunde däremot inte påvisas i en sådan omfattning att de skulle kunna förklara språkversionernas inekvivalens.

Slutsatser och diskussion

Alla fyra studier i denna avhandling berör aspekter av förhållandet mellan språk och matematik, närmare bestämt, möjliga effekter som språk kan ha på skriftliga matematikuppgifter. Det visade sig att språket på olika sätt kan ha inverkan på matematikuppgifter med konsekvenser för deras validitet. Dels kan vissa lingvistiska egenskaper vara möjliga källor för svårigheter att läsa och lösa uppgifter, dels kan matematikuppgifter översatta till olika språk vara inekvivalenta till följd av lingvistiska eller andra skillnader mellan språkversionerna. Några egenskaper

identifierades också som kunde kopplas både till inekvivalens och till svårigheter att läsa och lösa uppgiften. Om till exempel språkversion A använder sig av ovanligare ord än språkversion B, kan kravet på läsförmåga i version A öka. Det kan ha som följd att språkversion A blir svårare att lösa än språkversion B, versionerna är då inte ekvivalenta.
Resultaten i studierna i denna avhandling visar också att det inte är möjligt att ge några enkla lösningar för att undvika onödiga krav på läsförmåga eller inekvivalens mellan språkversioner. Att till exempel bara undvika långa ord eller långa meningar sänker inte nödvändigtvis kravet på läsförmåga. Att bara undvika skillnader i användning av aktiv och passiv form eller andra och tredje person mellan språkversioner höjer inte automatisk ekvivalensen. Istället är det nödvändigt att ta många olika faktorer i beaktande för att undvika både onödiga krav på läsförmåga och inekvivalens. Till exempel är inte bara längden av en mening avgörande. Även dess struktur och komplexitet kan spela roll för hur lätt det är för elever att förstå den, och då ha möjlighet att lösa matematikuppgiften korrekt.

Zusammenfassung auf Deutsch

Beim Lehren, Lernen und Betreiben von Mathematik spielt die Sprache eine wichtige Rolle. Mathematischer Inhalt muss mit Worten formuliert werden um kommuniziert werden zu können. Aber auch andere, nonverbale Formen der Kommunikation kommen zur Anwendung, wie zum Beispiel mathematische Symbole, Graphen oder Diagramme. Auch diese können als Teil der Sprache angesehen werden, die für mathematischen Diskurs verwendet wird. Das übergreifende Thema dieser Doktorarbeit ist die Untersuchung möglicher Effekte, die Sprache auf Mathematikaufgaben und damit auch auf die Validität von Studien zur Messung von Mathematikkompetenz haben kann. Zwei unterschiedliche Aspekte werden hier untersucht, um die zwei übergreifenden Fragen in dieser Doktorarbeit zu beantworten. Die erste dieser Fragen handelt davon, welche sprachlichen Eigenschaften einer Mathematikaufgabe, wie zum Beispiel die durchschnittliche Wort- oder Satzlänge, den Schwierigkeitsgrad beeinflussen, die Aufgabe zu lesen und/oder zu lösen. Zur Beantwortung dieser Frage werden in Studie I und II dieser Doktorarbeit Mathematikaufgaben auf Deutsch, Englisch und Schwedisch untersucht. Die zweite übergreifende Frage betrifft Äquivalenz verschiedener Sprachversionen einer Mathematikaufgabe, die Schülern in ihrer jeweiligen Muttersprache gestellt wird. Das Ziel ist es, mögliche Ursachen für Inäquivalenz der Sprachversionen zu ermitteln. Sowohl linguistische Unterschiede zwischen den Aufgabentexten als auch andere mögliche Faktoren, die die Äquivalenz beeinträchtigen können, werden untersucht. Diese Frage wird in Studie III und IV dieser Doktorarbeit behandelt, wo ebenfalls Mathematikaufgaben auf Deutsch, Englisch und Schwedisch untersucht werden. Detailliertere Forschungsfragen bezüglich dieser zwei Aspekte finden sich in den vier Studien dieser Doktorarbeit.

Der erste Aspekt: Linguistische Eigenschaften gekoppelt mit Schwierigkeiten Mathematikaufgaben zu lesen und zu lösen

Um eine schriftliche Mathematikaufgabe zu lösen, benötigt ein Schüler außer mathematischer Kompetenz immer auch ein gewisses Maß an Lesekompetenz, damit er den Aufgabentext lesen und verstehen kann. Der Text kann kurz sein und lediglich mathematische Symbole enthalten (z.B. $3 + 5 = ?$), aber auch in längeren Sätzen und Alltagssprache formuliert sein (z.B. „Amalia hat drei Meerschweinchen und fünf Kaninchen. Wie viele Tiere hat sie insgesamt?“). Da mathematische Kompetenz auch die Fertigkeit beinhaltet, Mathematik zu kommunizieren, kann man einen Teil dieser Lesekompetenz als Teil der Mathematikkompetenz ansehen. Unter anderem Wörter und Ausdrücke mit mathematischen Inhalt (z.B. *subtrahieren*, *Achsenabschnitt* oder *Integral*) sind Bestandteile einer mathema-

tischen Sprache, die für eine solche Kommunikation benötigt wird. Effekte anderer linguistischer Eigenschaften des Textes von Mathematikaufgaben, die Lesekompetenz erfordern, (wie z.B. ungewöhnliche, nicht-mathematische Wörter oder unnötig komplizierter Satzbau) können aber nicht zur mathematischen Kompetenz gezählt werden. Solche Eigenschaften gilt es, in Aufgaben, die zur Evaluierung der mathematischen Kompetenz gedacht sind, zu vermeiden. Ansonsten besteht die Gefahr, dass die Aufgabe zum Teil auch Lesekompetenz misst, die nicht Teil der Mathematikkompetenz ist.

Um den Zusammenhang zwischen einigen linguistischen Eigenschaften und der unnötigen Anforderung an Lesekompetenz (Demand of Reading Ability – DRA) zu untersuchen, wurden in Studie I Mathematikaufgaben der PISA-Studie 2012 auf Deutsch, Englisch und Schwedisch analysiert. Für jede Sprache wurde mit Hilfe einer Hauptkomponentenanalyse (Principal Component Analysis - PCA) der Grad jeder Aufgabe an DRA gemessen und mithilfe der Lösungsfrequenz auch der Schwierigkeitsgrad jeder Aufgabe bestimmt. Im nächsten Schritt wurden mögliche Korrelationen zwischen DRA bzw. Schwierigkeitsgrad und vier verschiedenen linguistischen Eigenschaften der Aufgabe untersucht, nämlich Wort-, Satz- und Textlänge sowie Informationsdichte. Die Resultate ergaben keinerlei Korrelationen für die Aufgaben auf Englisch. Auch für die Aufgaben auf Schwedisch konnten keine Korrelationen zwischen den linguistischen Eigenschaften und DRA gefunden werden. Die einzigen statistisch signifikanten Korrelationen fanden sich für die Aufgaben auf Deutsch und Schwedisch zwischen Schwierigkeitsgrad und Informationsdichte sowie für die deutschen Aufgaben auch zwischen DRA und sowohl Informationsdichte als auch Wortlänge.

Um mögliche Ursachen für DRA genauer zu untersuchen, haben in Studie II zwölf schwedische Schüler mit zwei der Aufgaben gearbeitet, die einen hohen DRA-Wert hatten. Während ihrer Arbeit mit den Aufgaben sollten die Schüler laut lesen und denken. Auf diese Weise konnten Variationen im Leseprozess der Schüler identifiziert werden, die Hinweise auf mögliche Ursachen für die erhöhte Anforderung an die Lesekompetenz in diesen Aufgaben darstellen können. Dies waren zum Beispiel ungewöhnliche, nicht-mathematische Wörter und komplexer Satzbau.

Der zweite Aspekt: Ursachen für Inäquivalenz zwischen unterschiedlichen Sprachversionen von Mathematikaufgaben

In manchen Situationen werden Mathematikaufgaben in verschiedenen Sprachen gestellt. Dies kann zum Beispiel bei landesweiten Untersuchungen von Mathematikkompetenz in mehrsprachigen Ländern der Fall sein. Aber auch für internationale Studien wie PISA (Programme for International Student Assessment) oder TIMSS (Trends in International Mathematics and Science Study)

müssen Aufgaben in viele verschieden Sprachen übersetzt werden. Um die Validität solcher internationaler Studien nicht zu gefährden, ist es wichtig, dass die unterschiedlichen Sprachversionen so äquivalent wie möglich sind. Das bedeutet vor allem, dass sie den gleichen Inhalt mit einer vergleichbaren Schwierigkeit behandeln, aber auch linguistische Äquivalenz wird angestrebt.

In Studie III wurde untersucht, ob Unterschiede bezüglich dreier linguistischer Eigenschaften (Anwendung von Aktiv/Passiv, Anwendung von zweiter/dritter Person, unterschiedliche Komplexität im Satzbau) einen Zusammenhang mit Mangel an Äquivalenz zwischen Sprachversionen von Mathematikaufgaben aufweisen, die Schülern in ihrer jeweiligen Muttersprache (Deutsch, Englisch oder Schwedisch) gestellt wurden. Um Inäquivalenz zwischen den Sprachversionen nachzuweisen, wurde eine Differential Item Functioning (DIF) Analyse durchgeführt. Für die DIF-Analyse wurden die drei Sprachversionen der Mathematikaufgaben der PISA-Studie 2012 paarweise untersucht. Weiterhin wurden Unterschiede zwischen den Sprachversionen bezüglich der drei oben genannten linguistischen Eigenschaften untersucht. Die Analysen ergaben, dass DIF bei einigen Aufgaben auftrat, das heißt die Sprachversionen waren dann nicht äquivalent. Am meisten trat DIF zwischen den englischen und schwedischen Versionen auf. Auch Unterschiede bezüglich der drei linguistischen Eigenschaften kamen vor, wobei der häufigste Unterschied der Gebrauch von Aktiv und Passiv war. Es konnten jedoch keine Korrelationen zwischen DIF und den Unterschieden gefunden werden. Ein möglicher Grund für dieses Ergebnis kann zum Beispiel sein, dass Schüler im Alter von 15 Jahren, für die die PISA-Aufgaben gedacht sind, eine so hohe Sprachkompetenz haben, dass die hier untersuchten linguistischen Eigenschaften deren Möglichkeit, die Aufgaben zu lösen, nicht wesentlich beeinflussen.

Um mögliche Ursachen für Inäquivalenz zwischen unterschiedlichen Sprachversionen von Mathematikaufgaben zu identifizieren, wurde in Studie IV eine qualitative Untersuchung durchgeführt. Interviews mit 26 Schülern aus Deutschland und 16 aus Schweden, die mit Aufgaben gearbeitet hatten, die in Studie III einen hohen Grad an DIF zwischen der deutschen und schwedischen Version aufgewiesen hatten, wurden qualitativ analysiert. In dieser Studie wurden außer Unterschieden bezüglich linguistischer Eigenschaften (Verwendung ungewöhnlicher Wörter, unterschiedliche Formulierung gewisser Sätze) auch andere mögliche Ursachen für Inäquivalenz zwischen den Sprachversionen identifiziert. Zum Beispiel reagierten die deutschen Schüler deutlich mehr auf Informationen im Aufgabentext, die nicht zum Lösen der Aufgabe nötig waren. Dies galt sowohl für erklärenden Text mit Hintergrundinformationen als auch für quantitative Angaben, die nicht zur Berechnung notwendig waren. Kulturelle Unterschiede zwischen den deutschen und schwedischen Schülern, die die Inäquivalenz erklären könnten, wurden in den Interviews aber nicht gefunden.

Schlussfolgerungen und Diskussion

Alle vier Studien in dieser Doktorarbeit behandeln Aspekte des Verhältnisses zwischen Mathematik und Sprache, genauer gesagt, mögliche Effekte die Sprache auf Mathematikaufgaben haben kann. Mehrere mögliche solcher Effekte konnten aufgezeigt werden. Zum einen können gewisse linguistische Eigenschaften mögliche Ursachen für Schwierigkeiten beim Lesen und Lösen von Mathematikaufgaben sein, während andere (z.B. Satzlänge) keinen solchen Effekt aufwiesen. Zum anderen können Mathematikaufgaben, die in verschiedene Sprachen übersetzt sind, aufgrund linguistischer und anderer Unterschiede Inäquivalenz aufweisen. Es wurden auch Eigenschaften identifiziert, die sowohl mögliche Lesehürden als auch Inäquivalenz mit sich führen können. Wenn zum Beispiel in Sprachversion A ungewöhnlichere Wörter als in Sprachversion B benutzt werden, kann die Anforderung an die Lesekompetenz in Version A höher sein. Dies kann zur Folge haben, dass Version A schwieriger zu lösen ist als Version B und die Versionen demzufolge inäquivalent sind.

Die Resultate in dieser Doktorarbeit haben auch gezeigt, dass es nicht möglich ist, einfache Anweisungen zur Verhinderung von unnötigen Anforderungen an Lesekompetenz oder Inäquivalenz zwischen Sprachversionen zu geben. Alleine die Vermeidung langer Wörter oder Sätze genügt nicht notwendigerweise, um Anforderungen an Lesekompetenz zu senken. Nur die Vermeidung von Unterschieden im Gebrauch von Aktiv und Passiv oder zweiter und dritter Person erhöht nicht zwangsläufig die Äquivalenz. Stattdessen müssen mehre unterschiedliche Faktoren beachtet werden, um unnötige Anforderungen an Lesekompetenz oder Inäquivalenz zu vermeiden. Zum Beispiel ist nicht nur die Länge eines Satzes entscheidend, sondern auch seine Struktur und Komplexität können Einfluss darauf haben, wie leicht Schüler ihn verstehen können und dadurch die Möglichkeit haben, die Mathematikaufgabe korrekt zu lösen.

1 Introduction

I have an interest in the relation between mathematics and language, which originates from my background as a mathematics teacher at schools in both Germany and Sweden and having taught mathematics in two different languages. By having learned mathematics in one language and later teaching mathematics in another language, I experienced how language and mathematics are intertwined, maybe more than teachers dealing with mathematics in just one language. In this thesis, I investigate some aspects of the relation between mathematics and language, with a particular focus on mathematics tasks.

1.1 Mathematics and language

Language plays an important role when teaching, learning and practicing mathematics (Morgan, Craig, Schuette, & Wagner, 2014). Mathematical content has to be put into words to be communicated between individuals, but also non-verbal forms of communication such as mathematical symbols, graphs, and diagrams are used to communicate mathematics and can be seen as a part of the language used in mathematical discourse.

To be able to practice mathematics, one has to be proficient in using language in an adequate way. Helping students to learn this use of language is a part of teaching mathematics, since mathematical communication is an important aspect of mastering mathematics (e.g., Lithner et al., 2010; Niss & Højgaard, 2011). Therefore, language has to be used in a way that aids students' mathematical thinking and understanding. At the same time, there is a risk that language forms an obstacle instead, for example, when mathematics tasks or instructions are formulated in a way that makes it hard for students to understand the meaning or content. The focus of this thesis is on aspects of language in relation to unnecessary difficulties of mathematics tasks given to students.

Mathematics tasks have been used for different purposes in different cultures for thousands of years (Johansson, 2004). Tasks have been used in education to make students practice but also to assess students' mathematical knowledge and ability. When a student is asked to solve a task, he or she also needs some linguistic ability, either purely oral if the task is given orally or some reading ability if the task is given in written form. When a written task is supposed to measure the mathematical ability of students, the validity of the results is jeopardized if the text of the task contains linguistic features that make it unnecessarily difficult for the students to read and understand the task. In this case, the task may to some extent assess students' reading ability instead. One focus of this thesis is therefore

the relation between linguistic features of written mathematics tasks and the students' difficulties in reading and/or solving the tasks.

Another aspect of language in relation to solving mathematics tasks this thesis focuses on, is *equivalence* of different language versions of mathematics tasks. Equivalence concerns both how similar the different language versions of tasks' texts are regarding formal features and also functional aspects, that is, how similar the different language versions are regarding content and difficulty. Equivalence is important in multilanguage assessments, that is, assessments including students of two or more language groups, since inequivalence of the tasks might jeopardize the validity of the results. Multilanguage assessments are, for example, used in nationwide tests in bi- or multilingual countries, for example, the Pan-Canadian Assessment Program (PCAP) (O'Grady, 2018) and in international assessments like the Programme for International Student Assessment (PISA) (OECD, 2014). One possible threat to equivalence of the language versions of tasks are linguistic differences between the task texts of the versions (Roth, Oliveri, Sandilands, Lyons-Thomas, & Ercikan, 2013). But other issues can also threaten the equivalence, like differences in cultural relevance of the context of the task (e.g., Yildirim & Berberoğlu, 2009) or curricular differences between the student groups (Huang, Wilson, & Wang, 2014) and by that imperil the validity of the results.

In this thesis, I investigate these two aspects of the relation between language and mathematics in connection with mathematics tasks in written form.

1.2 Purpose of the thesis

The purpose of this thesis is to investigate some of the effects that language can have on written mathematics tasks, in particular, on the validity of mathematics assessments. This thesis focuses on effects in relation to demands of reading ability and in relation to equivalence in multilanguage assessments. This is done by investigating the following two overarching questions:

- Which linguistic features of mathematics tasks are related to difficulties in reading and/or solving the tasks?
- Which are possible sources of inequivalence between different language versions of mathematics tasks?

To answer these questions, both quantitative and qualitative methods are used to study tasks given to students in their instructional language English, German, or Swedish. Each overarching question is addressed by two of the studies included in this thesis, one quantitative and one qualitative study for each question. An

overview of how the four studies included in this thesis are connected to each other is given in figure 1.

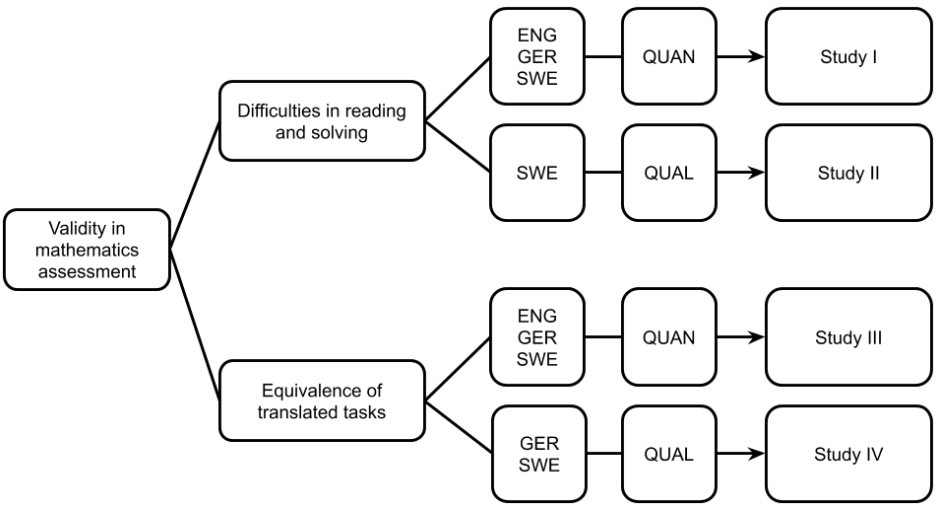

Figure 1 Overview of the four studies included in this thesis.

2 Background

The major theme of this thesis is mathematics and language in relation to the validity of mathematics tasks. In addition, there are two overarching questions, one regarding linguistic features related to difficulties in reading and/or solving a task and the other regarding issues of equivalence of translated mathematics tasks. In the following sections, I will first give an overview about and discuss the relation between mathematics and language in general terms. In connection to the first overarching question of this thesis, a section about the demand of reading ability of a mathematics task follows. After that, connected to the second overarching question, a section follows that concerns equivalence and sources of inequivalence of mathematics multilanguage assessment and ways to measure inequivalence statistically.

2.1 Mathematical language

There are different views about how language and mathematics relate to each other. One view is that mathematics *is a language* by itself. For example, Kaiser and Schwarz (2003, p. 374) include in their concept of mathematical literacy the standpoint that mathematics is a language that students have to learn. Wakefield (2000) gives several examples of the similarity of properties between a language and mathematics, such as the use of uniform and consistent symbols and the need of encoding and decoding to communicate. She also sees the fact that the meaning is influenced by the order of symbols, that is, the existence of a syntax, as a similarity between mathematics and language. Therefore, Wakefield also takes the standpoint that mathematics is a language that students have to learn like a second language. A similar standpoint is taken by (Bullock, 1994, p. 736), who argues that if mathematics had only consisted of new words and symbols, than it could have been seen as an extension of existing natural language. But since it also has its own syntax and grammar, he sees it as an independent language. This language "math" is used for communication in situations where the natural language by itself is insufficient.

On the other hand, there are also arguments for the view that mathematics is not a language by itself but *has a language* that is used as a means to communicate mathematics. Learning this "mathematical language" is seen as a part of learning mathematics (Riccomini, Smith, Hughes, & Fries, 2015), and language cannot be separated from learning (Schleppegrell, 2007). This mathematical language has several characteristics, like special technical vocabulary (Riccomini et al., 2015), the use of certain grammatical patterns (Schleppegrell, 2007), and the use of formulas, graphs and other semiotic resources (Dyrvold, 2016). That is, a special

kind of "mathematical reading ability" is needed to master the mathematical language (Shanahan & Shanahan, 2008). Being able to use this mathematical language is necessary for communication of mathematics, and this ability can be seen as part of mathematical proficiency (see e.g., Lithner et al., 2010; NCTM, 2000; NRC, 2001). Still, the context in which mathematics is performed often has the need of natural, "non-mathematical" language, for example, when mathematical models are applied on real world problems.

In this thesis, I take this second standpoint, that a special mathematical language is used to do and communicate mathematics. This view on the relation between mathematics and language corresponds to the model in section 2.2 where non-mathematical reading ability is separated from reading ability related to mathematics.

2.2 Reading mathematics tasks

When mathematical content is communicated in written form, some reading ability is demanded of the individuals involved in the communication. This is the case for all different types of mathematical texts, like instructional texts in mathematics textbooks, exercise items, and tasks in assessments. When, for example, a mathematics task is given to students in written form in an assessment, they have to read and understand at least parts of the task's text to be able to solve the task. Thus, they need some reading ability to be successful. Depending on the task text, not only a different *amount* of reading ability is needed, but also different *kinds* of reading ability are demanded to understand the task text. The different kinds of a text's reading demand are presented in the following subsections.

2.2.1 Necessary reading demands

One part of the reading ability that is needed to be able to solve a mathematics task, is the ability to understand and make meaning of the mathematical language. This ability includes understanding the mathematical vocabulary, both technical terms that are specific for mathematics (e.g., *diameter*, *quadrilateral*, or *median*) and words that have a different or more specific meaning in mathematics than in everyday language (e.g., *product*, *function*, or *kite*) (Riccomini et al., 2015). It also includes being able to interpret grammatical constructions that represent mathematical relations (e.g., *less than or equal to*), decoding long, dense noun phrases (e.g., *the volume of a right circular cylinder*), and using numbers, graphs, and formulas (Schleppegrell, 2007). This component of reading ability, the ability to make meaning of written mathematical content, can be seen as a part of mathematics proficiency. That is, a task intended to assess mathematical ability can also include demands of this part of reading ability and still be valid. In figure 2, this mathematical reading ability is represented by the overlap between mathematics ability and reading ability, that is, section B.

2.2.2 Unnecessary reading demands

There can also a part of reading ability required to understand the text of a mathematics task, which is not part of mathematics proficiency. This part of reading ability concerns reading demands that are not connected to mathematics, that is, unnecessary reading demands. Ideally, the task text is formulated in a way that such demands of non-mathematical reading ability is minimized, by, for example, using appropriate and relevant language that is easy to understand for the students that the task is designed for. However, the unnecessary reading demands of a task can also be enhanced by features of the text like uncommon non-mathematics words, complicated grammatical structures, or other features decreasing the readability of texts. In figure 2, reading ability needed to solve a task due to unnecessary reading demands is represented by section C.

In mathematics tasks, unnecessary reading demands should be avoided as far as possible for different reasons. When, for example, a task is given to students for instructional or learning purposes, a high unnecessary reading demand is counterproductive, since it may shift the focus of the task from the mathematical content to reading problems. If the purpose of a task is to assess students' mathematical proficiency, the validity of the assessment is jeopardized by high unnecessary reading demands, because a task then may to some degree measure reading ability instead of mathematical ability due to unnecessary linguistic complexity (Abedi, 2006). Prediger, Wilhelm, Büchter, Gürsoy, and Benholz (2015) give examples where reading demands such as complex sentence structures were obstacles for students when solving mathematics tasks.

The model of the relation between mathematical ability and reading ability as shown in figure 2 is, of course, simplified, since the relation between the two abilities is complex. For example, Chen and Chalhoub-Deville (2016) found that the strength of the relation between reading language proficiency and mathematical performance differed depending on the mathematical ability of students. However, for the purpose of identifying linguistic features of task texts related to unnecessary reading demands, this model is adequate since it allows to separate the two parts of reading ability needed to solve a mathematics tasks as described in section 3.2.1

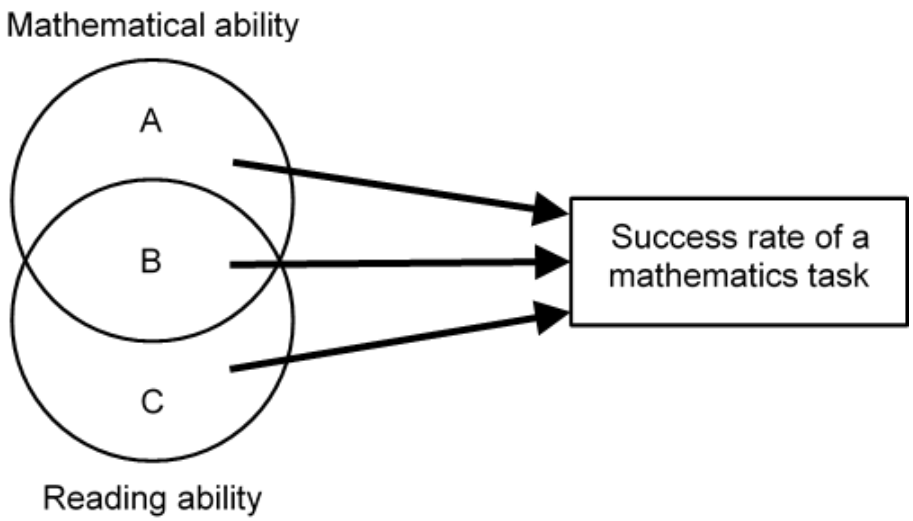

Figure 2 Schematic illustration of a theoretical model of the relation between mathematical ability and reading ability (Bergqvist, Theens, & Österholm, 2018).

In summary, when mathematics tasks are used to assess students' mathematical ability, tasks with high unnecessary reading demands are a threat to the validity of the assessment, since the tasks then may assess reading ability instead of mathematics ability to some degree. Therefore, such reading demands have to be avoided as far as possible.

2.2.3 Measuring unnecessary reading demands using Demand of Reading Ability (DRA)

To be able to measure unnecessary reading demands of a mathematics task, it is important to separate the necessary reading demands (as described in section 2.2.1) from the unnecessary reading demands (as described in section 2.2.2). Many methods to measure reading demands of a mathematics task cannot make this distinction. For example, when measuring the reading demand of a mathematics task by using the correlation of student results on this task with the students' results on reading tasks (e.g, Roe & Taube, 2006), only an overall reading demand of the task can be measured.

When Österholm and Bergqvist (2012a) compared different methods to measure reading demands of mathematics tasks, it was shown that using a principal component analysis (PCA) is a reliable and valid method to separate and measure unnecessary reading demands. To do this separation, this method uses students' results on a set of mathematics tasks and a set of tasks that assess reading ability.

The PCA then gives a quantitative measure for the unnecessary reading demand of a mathematics tasks that is labeled as *Demand of Reading Ability* (DRA). The method is described in more detail in section 3.2.1.

To be able to minimize DRA in a mathematics task, it is necessary to know possible sources of unnecessary reading demands. Linguistic features that in earlier research showed to be related to DRA for tasks in Swedish were, for example, word length and information density (Österholm & Bergqvist, 2012b) and commonness of words (Dyrvold, Bergqvist, & Österholm, 2015).

2.3 Equivalence of tasks in different languages

Another threat to the validity of an assessment may arise in multilanguage assessments, that is, assessments given to students in different languages. When one language version of a task has higher reading demands than another language version of the same task, the different task versions measure mathematical and reading ability to different degrees and, by that, the results of the assessment are less comparable between the different language versions. The occurrence of different reading demands is only one of several possible threats to the equivalence of tasks translated to different languages. The following sections concern the equivalence of translated tasks in general, in relation to both linguistic features and other features of the assessment.

2.3.1 Concepts of equivalence

When tasks are translated to different languages, it is important that the different language versions are as equivalent as possible. In everyday language, equivalence can be interpreted as that the versions are "the same task", just written in different languages. But since the task text changes when it gets translated, the different language versions are never exactly "the same". For example, different languages have different inherent properties that often make literal translations incorrect. Still, the language versions of the task should be as equivalent as possible.

There are many different views on the concept of equivalence in multilingual settings. For example, T. P. Johnson (2006) identified 60 different forms of equivalence regarding cross-cultural-measurement in literature. In addition, Arffman (2010) distinguishes between equivalence in the view of translation theory and in the view of test theory. Equivalence in the view of *translation theory* means that the translated text and the source text share some 'sameness' (Panou, 2013), which includes several aspects of equivalence. To be equivalent in the view of translation theory, the different language versions therefore have to be equivalent regarding, for example, content, connotation, style and register, and formal-aesthetic aspects (Koller, 2011). In the view of *test theory*, equivalence of tasks means

that the different versions of a task measure the same concepts and assess the same skills with the same degree of difficulty (Arffman, 2010). Otherwise, the validity and reliability of the test are jeopardized. Similarly, (Grisay, De Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007, p. 263) define "equivalence in an international test [...] as an equal probability of getting any particular item correct for all students at a given level of proficiency, independent of the national version they were administered", that is, they focus on the test theory view of equivalence.

Since this thesis is located in the context of international multilanguage assessments of mathematical competence, the test theory view of equivalence is the dominating one. To minimize the risk of bias in this context, it is important that the different language versions given to the students in the different language groups are equal in both difficulty and content (Peña, 2007). However, equivalence in the translation theory view also plays a role in this thesis, since one of the overarching questions addresses the impact of differences in linguistic features on the equivalence regarding difficulty between the different language versions.

There is a manifold of threats to equivalence of different language versions of mathematics tasks. There can be linguistic differences between the language versions, resulting in, for example, a higher degree of unnecessary reading demands in one of the language versions, but also other sources of inequivalence may occur. In the following subsections, first linguistic and then other types of potential sources of inequivalence are described and exemplified.

2.3.2 Linguistic sources of inequivalence

In multilanguage assessment, *linguistic sources* of inequivalence are differences in linguistic features between the different language versions of the task that lead to inequivalence. Linguistic differences are, for example, differences in how complex the sentences are or differences in the use of uncommon words. To decrease the risk of inequivalence due to linguistic differences, complicating or simplifying the vocabulary and the syntax should be avoided when translating the task (OECD, 2010). Still, it is possible that the task's difficulty changes due to inherent properties of the languages involved when a mathematics task is translated to another language. This can, for example, happen due to differences in how easily mathematical expressions can be understood. For example, the mandarin term for "median" is 中位數 (Zhōng wèi shù) meaning literally "the number in the middle" (D. Zhou, personal communication, October 17, 2018). That is, it is easier to grasp the meaning of median for a Chinese student than for an English-speaking student. Another example is the Swedish word for "quadrilateral", which is "fyrhörning", meaning literally "four-cornered [figure]". This makes the word easier to understand in Swedish than in English, possibly giving an advantage to Swedish speaking students. A concrete example where this difference occurred is

a task from the Swedish national mathematics assessment for 9th graders in 2013 (PRIM-gruppen, 2019). A task supposed to assess students' knowledge in probability theory asked for the probability to draw a card showing a quadrilateral out of five cards showing different geometric shapes (2 out of 5 were quadrilaterals). In the Swedish version, the word "fyrhörning" was used, that is, even without any geometrical knowledge, a student can understand what is asked for and solve the task. To solve the English version of the task, given to classes with English as instructional language, some knowledge about the names of geometric shapes is also required, to be able to understand the meaning of "quadrilateral" and to solve the task.

Other linguistic features that often are seen as indicators of a text's readability are word and sentence length (e.g., Abedi, Leon, Wolf, & Farnsworth, 2008; Lenzner, 2014). If these features differ between different language versions of a task, this might be a source of inequivalence. On the other hand, changes in word or sentence length may sometimes be inevitable due to inherent properties of the languages when a task is translated. For example, compounds consisting of several words in English (e.g., "steam engine") are in German and Swedish generally written as one word ("Dampfmaschine" and "ångmaskin", respectively). Because of this, the words in the text get longer in average, but the sentences get shorter, if measured in words per sentence.

Other linguistic differences that might cause inequivalence in multilanguage assessment, can arise due to changes made during translation. In the translation and adaptation guidelines for PISA 2012 (OECD, 2010), for example, translators are advised not to change passive voice in the source text into active voice in the target text or vice versa in order to avoid inequivalence. The use of more or less of common or technical/scientific vocabulary in the different language versions may also be a source of inequivalence and should be avoided (OECD, 2010).

However, if two language versions share all linguistic features and are equivalent in the view of translation theory, they can still be inequivalent regarding reading demands. If, for example, both language versions use passive voice but the use of passive voice is less common in language A and more common in language B, there might be an advantage for the students taking the assessment in language version B. Thus, the task versions are not equivalent in the test theory view, because the task is more difficult to understand (and to solve) in language A.

In summary, there are many possible ways how equivalence between different language versions of mathematics tasks can be threatened by linguistic issues. Different types of linguistic differences have been identified as sources of inequivalence, such as differences in word difficulty (e.g., Allalouf, 2003; Ercikan et al.,

2010), grammatical and semantical differences e.g., Roth et al. (2013), or differences in text length (e.g.,Ercikan, Gierl, McCreith, Puhan, & Koh, 2004). However, other differences between the tasks may also reduce equivalence, which will be discussed in the next section.

2.3.3 Other sources of inequivalence

When tasks are translated for the use in multianguage assessments, not only linguistic differences are a threat to equivalence. Cultural differences between the student groups who use the different language versions can also be a source of inequivalence. An example is the PISA science assessment 2006, where Çikrikçi Demirtaşli and Ulutaş (2015) identified a task about the Grand Canyon (task S426) as biased in favor of American students compared to Turkish students, due to the cultural difference of familiarity with the content.

Another possible source of inequivalence is curricular differences between the groups of examinees. These differences include, for example, differences in how much a topic was covered by the curriculum for students of relevant age. Other possible curricular differences may be which aid (protractors, pocket calculators, etc.) or which types of tasks (open answer tasks, multiple choice, etc.) the students are used to. For example, Huang et al. (2014) found curricular differences as sources of inequivalence in the PISA science assessment 2006, when comparing Mainland Chinese and American students from Georgia. An example given in that study was that the Chinese students were due to curricular reasons expected to have a higher level of understanding of the necessary conditions needed for combustion than their American counterparts.

2.3.4 Measuring inequality in multilanguage assessment using differential item functioning (DIF)

A common way to identify tasks for which language versions are not equivalent is to use a statistical measure called differential item functioning (DIF). In general, an item displays DIF if participants with the same level of the measured attribute but belonging to different subgroups have different probability to give a certain answer (AERA, 2014, p. 51). The different subgroups can be characterized by, for example, gender, socio-economic background, language, or place of residence. To give a concrete example, imagine a nationwide test in geography in the USA. If one question would be “What is the capital of Idaho?” students living in Idaho would have higher probability to answer the question correctly compared to students from other states, even if Idaho students are not necessarily better in geography in general. That is, this question would display DIF in favor of students

from Idaho.[1] In the case of mathematics tasks in different languages, a task displays DIF if students with the same general mathematical ability but from different language groups have different probabilities to solve that particular task correctly. DIF can be detected by statistical methods. In this thesis, the Mantel-Hanszel-method was used for the DIF-analysis. It is described in more detail in the methods section.

To be able to avoid inequivalence between the different language versions of a translated task, sources of DIF have to be identified. This is mostly done using qualitative methods like reviews of the language versions by experts in the languages, culture, or curriculum (e.g., Allalouf, 2003; Ercikan et al., 2010; Roth et al., 2013). Other studies also used students' comments and impressions of the tasks in combination with expert judgments to identify sources of DIF (e.g., Benítez & Padilla, 2014; Ercikan et al., 2010). In this thesis, task-based interviews with students that had worked with DIF tasks were used to identify possible sources of DIF. The analysis is described in more detail in the methods section.

[1] Just for the records: The capital of Idaho is Boise.

3 Methods

Since the main theme of this thesis concerns effects that language can have on mathematics tasks, all four studies included are dealing with different aspects of the role language has in written mathematics tasks. To be able to identify linguistic features of mathematics tasks that are related to difficulties in reading and/or solving mathematics tasks and to find sources of inequivalence between different language versions of mathematics tasks, different methods have been used.

To detect *the occurrence* of unnecessary demand of reading ability (DRA) in a mathematics task or of inequivalence between language versions of tasks (i.e., occurrence of DIF), appropriate quantitative methods are used in Study I and Study III, respectively. Other quantitative methods were used in these two studies to identify possible *sources* of the phenomena that were measured. Still, these methods can only focus on a limited range of features in the tasks that may be such sources. Therefore, qualitative methods with a wider approach are used to aid the identification of features in the tasks that are possible sources of unnecessary reading demands in a task (Study II) or inequivalence between language versions (Study IV).

In this thesis, thus, quantitative and qualitative research methods, techniques, and approaches are combined into one mixed methods design. This sequential design, combining quantitative methods followed by qualitative methods, enables finding answers to the overarching questions in this thesis and the research questions of the studies included. A similar design was, for example, used by Benítez and Padilla (2014) when they investigated possible sources of DIF between the English (USA) and Spanish (Spain) versions of the PISA student questionnaire. After having identified DIF items quantitatively, the authors then could identify causes of DIF qualitatively through cognitive interviews with students.

By this approach, this thesis takes advantage of several of the positive effects of mixed methods design as described by R. B. Johnson and Onwuegbuzie (2004) and Cohen, Manion, and Morrison (2013). By using a two-stage sequential design, the stage 1 results (i.e., identification of occurrence of DRA or DIF) are used to develop and inform the purpose and design of the stage 2 component (i.e., finding sources of DRA or DIF). Mixed methods can also provide stronger evidence for a conclusion through convergence and corroboration of findings as, for example, the findings in this thesis regarding passive voice (see section 4.2). Further, the combination of quantitative and qualitative methods can add insights and understanding that might be missed when only a single method is used.

This section presents the data used in the studies, how the two main phenomena (unnecessary reading demands and inequivalence) were localized, and how the sources of these phenomena were examined. Lastly, the ethical considerations are described. An overview about the data and the methods used in the studies is given in Table 1.

Table 1: Overview of the studies in this thesis, their data and methods.

Study	Objective	Data	Method
Study I	Investigation of the relation between linguistic features of mathematics tasks and difficulties in reading and solving the tasks.	Results of PISA 2012 tasks in mathematics and reading in ENG (USA), GER, SWE. Text of PISA mathematics tasks.	Statistical methods, correlations. (quantitative)
Study II	Investigating variations in students' reading process to identify sources of difficulties in reading mathematics tasks.	Think-aloud-protocols of 12 students working with SWE mathematics tasks of PISA 2012.	Analysis of the reading processes. (qualitative)
Study III	Investigation of inequivalence and differences in linguistic features between different language versions of mathematics tasks. Identify possible sources of inequivalence.	Results and text of PISA 2012 tasks in mathematics in ENG (USA), GER, SWE.	Statistical methods to measure inequivalence and linguistic features, correlations. (quantitative)
Study IV	Identify possible sources of inequivalence between language versions of mathematics tasks.	PISA mathematics tasks, task-based interviews of 16 Swedish and 26 German students.	Analysis of issues and difficulties the students encounter. (qualitative)

3.1 Data used in the studies

All four studies in this thesis used data obtained from tasks of the Programme for International Student Assessment (PISA) 2012. For the quantitative studies, student results from the PISA assessment 2012 (OECD, 2012a) in mathematics and reading and data derived from the original PISA 2012 mathematics task texts were used. For the qualitative studies, data were obtained from students that

worked with some of the mathematics PISA tasks. The data are described in more detail in the different subsections below.

There were several reasons to use the PISA tasks from 2012. Firstly, the tasks are professionally translated to different languages according to a rigorous procedure (OECD, 2010). Therefore, it can be assumed that pure translation errors are avoided as far as possible. Secondly, the tasks were solved by thousands of pupils (OECD, 2014) in different countries. That is, there is a big data material available, and the anonymized PISA results for every student are freely accessible at the OECD website (OECD, 2012a). Thirdly, there are results both for tasks assessing the students' mathematical ability and their reading ability, which makes it possible to calculate a measure for the unnecessary demand of reading ability of a task (see Section 3.2.1).

The drawback of using PISA tasks is the fact that most of the tasks are confidential. Therefore, it was necessary to apply for permission by the OECD and the national PISA authorities to use the tasks in the studies. Although the permission was granted, it is not allowed to publish the confidential tasks used in the studies. Therefore only one (non-confidential) task could be given as an example in the appendix of Study IV.

3.2 Occurrence of unnecessary reading demands and inequivalence

To identify the presence of unnecessary reading demands in a mathematics task and inequivalence between language versions of a task, different quantitative measures are suitable: demand of reading ability (DRA) and differential item functioning (DIF), respectively. The methods used for the identification and measurement of DRA and DIF are described in the sections below.

3.2.1 Measurement of demand of reading ability

As described in the background, unnecessary demand of reading ability should be avoided in mathematics tasks. To be able to do this, it is necessary to first identify the occurrence of DRA. For this purpose, I chose a method using a Principal Component Analysis (PCA). This method has shown to be able to measure a task's DRA in a valid and highly reliable manner and is superior to other possible methods like, for example, regression analysis (Österholm & Bergqvist, 2012a). The PCA was performed on the data of the PISA results from the OECD (2012a) using the statistical software SPSS with suitable extensions. To further test the reliability of this method, I also performed the PCA on a randomly chosen selection of only half of the students' results and got similar results as in the analysis of all data. This procedure indicates that the method is reliable.

In short, the measurement of DRA using the PCA was done like this: The PCA was performed on a dataset consisting of students' results both on PISA tasks intended to measure mathematics literacy and tasks intended to measure reading literacy. The mathematics tasks were expected to load mainly on one component (corresponding to mathematical ability) and the reading tasks on another component (corresponding to reading ability). However, some mathematics tasks also had a high loading on the "reading component", which gave a measure for the unnecessary demand of reading ability in the mathematics task.

3.2.2 Identifying inequivalence

To identify inequivalence between language versions of tasks, an analysis of Differential Item Functioning (DIF) is performed. To quantitatively identify tasks that display DIF when given to different groups of participants several different methods can be used. For example, Yildirim and Berberoğlu (2009) used methods based on Factor Analysis, based on Item Response Theory (IRT), and the Chi-Square based Mantel-Haenszel method to detect DIF between the English and Turkish version of mathematics task of PISA 2003. The authors found high agreement rates between the different methods. Also, Hambleton and Rogers (1989) got similar results when using IRT and Mantel-Haenszel to identify DIF in the New Mexico High School Proficiency Exam between groups of Anglo-American and Native American students.

In this thesis, I decided to use the Mantel-Haenszel method to calculate DIF because it is one of the most commonly used methods for DIF-detection (Dorans & Holland, 1992; Michaelides, 2008). The method is based on the calculation of odds-ratios and was possible to perform on the PISA results from the OECD (2012a) without major adaptations. The results of the analysis of different datasets were replicable. An alternative would have been a DIF-detection method based on IRT. This would have required more complicated modelling and calculations, but since the Mantel-Haenszel method gives similar results (Hambleton & Rogers, 1989), I decided on using the more simple Mantel-Haenszel method.

3.3 Sources of unnecessary reading demands and inequivalence

To identify possible sources of unnecessary reading demands in a mathematics task or potential causes of inequivalence between language versions of a task, different methods are suitable. When *quantitative* statistical methods are used to identify these sources, only a limited number of features of the tasks can be investigated. Therefore, also *qualitative* research methods with the possibility for a broader and more open approach were used in the studies of this thesis.

In subsection 3.3.1, the quantitative methods that were used in study I and III to identify possible sources of unnecessary reading demands (measured by DRA) and inequivalence (measured by DIF) are presented. The qualitative methods used in study II and IV are presented in the following subsections.

3.3.1 Quantitative methods to identify sources of unnecessary reading demands and inequivalence

To identify possible sources of unnecessary reading demands and inequivalence, Study I and III in this thesis investigated quantitatively the correlation between a range of linguistic features of the texts of mathematics tasks and DRA and DIF, respectively. The selection of linguistic features for investigation in these two studies was guided by two main criteria. Firstly, features were chosen that had shown to be sources of reading difficulties or difficulties in solving mathematics tasks in earlier research. Secondly, it had to be possible to operationalize these features to be able to calculate correlations. This operationalization was done by, for example, calculating the average number of words per sentence or the number of occasions for use of passive voice for each task. Then, correlation coefficients for the relation between the values for the linguistic features and the measures for DRA or DIF could be calculated.

3.3.2 Qualitative methods to identify sources of unnecessary reading demands and inequivalence

The main advantage of *qualitative* research methods compared to quantitative methods is the possibility for a broader and more open approach. The use of qualitative methods allows to analyze the data in a more exploratory way, making it possible to identify a broad variety of different types of possible sources of unnecessary reading demands and inequivalence. For example, issues that are experienced as problematic by students when working with the tasks may be discovered, that were not expected beforehand and would not have been identified with quantitative methods.

The data used for the qualitative studies in this thesis were think-aloud-protocols (TAP) of students working individually with some of the mathematics PISA 2012 tasks (Study II) and task-based interviews of the students after their work with these tasks (Study IV). The tasks given to the students were tasks that were identified as displaying DIF between the German and the Swedish versions in study III. Two of these tasks had also shown high DRA in the Swedish version in Study I and the TAPs concerning these two tasks were analyzed in Study II. The data collection was conducted in spring 2018 in 3 Swedish and 4 German schools.

3.3.3 Selection of schools and participants

The selection of participants in both Sweden and Germany was guided by the aim to get a varied sample of students in both language groups, despite the practical limitation regarding the number of participants. That is, the participation of both girls and boys with different social background, and both higher and lower attainment students in both language groups was desirable. Therefore, in both Germany and Sweden, the students were recruited from several different schools from both rural and urban areas in different municipalities. The German students came from the German state of Schleswig-Holstein, where students choose different school types after year four, depending on their level of educational attainment. To get a diverse selection of students, both three schools for students with higher attainment ("Gymnasium") and one for students with lower attainment ("Gemeinschaftsschule") were chosen for this study. In Sweden, all students attend the same school type until year nine, that is, this kind of selection was not necessary there, but diversity was achieved by including schools from different areas

Having very low attainment students or not native speakers as participants, could shift the focus of this study from more common issues possibly being sources of inequivalence towards specific problems these students encounter. Therefore, the students' teachers aided the selection of suitable participants to avoid students with these traits in the studies.

The final selection of participants consisted of 16 Swedish and 26 German students, between four and nine students from each school. The difference in the number of students in the language groups arose because there were more German students than expected, who wanted to join the studies. Since the quality of the studies is not jeopardized by having more participants in one of the language groups, all were allowed to participate. All students were about the age of 15, which is the age for which the PISA tasks are designed.

3.3.4 Think-aloud-protocols

To enable identification of words, phrases, or sentences that were problematic for the students in some way, the students were asked to read and think aloud while they were working with the tasks allotted to them. That is, a more precise designation would be read-and-think-aloud-protocol, but I chose to use the more common name think-aloud-protocol (TAP) in this thesis.

Reading aloud when working with a mathematics task is a somewhat unnatural way for the students to tackle a mathematics task. Still, difficulties when reading aloud can indicate general reading problems with a text since oral reading fluency is an indicator for reading competence in general (Fuchs, Fuchs, Hosp, &

Jenkins, 2001). Similarly, there could also have been a risk that thinking aloud when working with a task might influence and change the students' thinking processes. However, Ericsson and Simon (1993) conducted a review of empirical studies on the effect of verbalizations on the thought process and found no evidence for such an effect. Thus, it is unlikely that the students' thought processes are altered by the verbalization of their thoughts, while attempting to solve a task. Therefore, TAPs are a suitable mean to get insights about obstacles students encounter when working with mathematics tasks.

To find possible sources of unnecessary reading demands, the TAPs of Swedish students during their work with Swedish tasks that had high DRA were analyzed in Study II. In Study IV, the think-aloud-phases served as base for task-based interviews, used to identify possible sources of inequivalence, as described in the following section.

3.3.5 Task-based interviews

To enable the identification of sources of inequivalence, the students were interviewed after their work with each task. They were asked about how they solved the task and about their perceptions of difficulties. The interviews were conducted with an interview guide containing both open questions to enable a broad approach on experienced difficulties and questions concerning different concrete types of issues that might be problematic. For example, questions were asked about linguistic difficulties (e.g., whether the students experienced any words or sentences as difficult) or curricular issues (e.g., whether the students had worked with similar tasks before) since these types of issues often were identified as sources of DIF in earlier research (e.g., Allalouf, Hambleton, & Sireci, 1999; Huang et al., 2014). For each task, there were also some task-specific questions regarding issues unique for the task that were expected to possibly be problematic in some way (e.g., the use of Fahrenheit).

During the interviews, the interview guide was used flexibly as, for example, advocated by Goldin (2000). This flexibility means that depending on the students' answers, some questions could be taken in different order. Still, not to jeopardize reproducibility, the interviews followed the same path, so that all students got "the same interview" with the same questions and only minor deviations.

3.4 Ethical considerations

When involving human beings in scientific studies, it is necessary to ensure that the participants are not harmed in any way. Therefore, possible ethical issues have to be taken into consideration. Throughout the studies in this thesis, the guidelines stated by the Swedish Research Council for research in social science (Vetenskapsrådet, 2011) were followed.

Since it is impossible to identify any student in the OECD data, and the students participating in Study II and IV are not the same that worked with the PISA tasks in 2012, no ethical issues occur in Study I and III.

When interviewing the students in Study II and IV, ethical issues had to be considered. Although no highly sensitive information was collected, it is always possible that some personal information gets revealed during interviews and TAPs. Therefore, informed consent had to be obtained from the students, that is, they got information about the purpose of the study and how data gets collected handled. They were also informed that their participation is voluntary and that they can withdraw from the study whenever they want without giving any reasons. According to the national regulations in Germany, the students' parents had also to give their consent and a permission to conduct the study at schools had to be obtained from the school ministry of the state of Schleswig-Holstein.

All collected data were anonymized, making it impossible to identify any of the students afterwards.

4 Results

The results of the four studies included in this thesis are presented in the following sections. Section 4.1 deals with the first overarching question in this thesis, that is, *which linguistic features of mathematics tasks are related to difficulties in reading and/or solving the tasks?* Answers to this question are given in Study I & II by first measuring the amount of unnecessary demand of reading ability (DRA) and the difficulty of solving the tasks and then identifying features of the task texts that may be sources of DRA or difficulty.

In section 4.2, the second overarching question is dealt with: *Which are possible sources of inequivalence between different language versions of mathematics tasks?* Answers to this question are given in Study III & IV, where tasks with inequivalence between language versions are identified and then analyses are conducted to find possible sources of inequivalence.

4.1 Linguistic features related to difficulties in reading and solving

The first general question in this thesis concerns the linguistic features of mathematics tasks that are related to difficulties in reading and solving the tasks. This question is answered with results from the first two studies included in this thesis. Below, I first present results from each of the studies for themselves and then a combination of the results.

The relation between four linguistic features of the texts of mathematics tasks of PISA 2012 in English, German, and Swedish and the tasks' difficulties in reading (measured by DRA) and difficulties in solving (measured by success rate) was investigated quantitatively in Study I. The linguistic features investigated in this study were word length, sentence length, task length, and information density.

The measures for word length and information density were significantly correlated with DRA for the German tasks, but for the task versions in English and Swedish, no significant correlations between the linguistic features and DRA were found. Also, for the German tasks, more of the variance of DRA could be explained by the linguistic features, compared to English and Swedish.

Regarding difficulties in solving, only information density was significantly correlated to success rate and only for the tasks in German and Swedish. But also here, the linguistic features investigated could explain more of the variance for the German tasks, compared to English and Swedish.

In conclusion, the German tasks showed to be more affected by the linguistic features investigated in Study I than the tasks in English and Swedish, both regarding difficulty in reading and difficulty in solving.

Still, tasks with high DRA were identified in all three languages in Study I but there were no correlations between the linguistic features investigated and DRA for the Swedish and English tasks. Therefore, there has to be other reasons for the occurrence of unnecessary reading demands in these tasks.

To identify features of the Swedish tasks possibly related to unnecessary reading demands, the variations in the reading process of Swedish students working with two of the tasks showing high DRA were investigated in Study II by analyzing think-aloud-protocols. These variations could indicate linguistic features possibly increasing reading difficulty. Common variations on word-level were, for example, that students were hesitating or stumbling when they read some words. Also misreading of words occurred. On sentence-level, rereading of certain sentences was a common variation. With these findings, some uncommon words (e.g., *criterion*) and complex sentences were identified in the task texts that may be obstacles for students when reading and solving the tasks in Swedish and by that be possible sources of unnecessary reading demands.

The findings of the quantitative analyses in Study I did not only enable the qualitative analysis in Study II, but the combination in a mixed-methods design also gave more information about some linguistic features that are possible sources of unnecessary reading demands. Word length and sentence length were not related to difficulties in reading in Study I, but still, some words and sentences appeared to be reading obstacles in Study II. Therefore, other properties of the words and sentences than purely length are probably reasons why these words and sentences make it more difficult to read the task text. The results of study II indicated that uncommonness of words and complexity of sentences may be these properties.

4.2 Sources of inequivalence in multilanguage assessment

The second general question in this thesis concerns possible sources of inequivalence between different language versions of mathematics tasks. This question is answered with results from the third and fourth study included in this thesis. As in the preceding section, I first present results from each of the studies for themselves and then a combination of the results to answer the second general question.

To find possible sources of inequivalence between different language versions of mathematics tasks, in study III, the English, German, and Swedish versions of

mathematics PISA tasks were analyzed in several steps. To measure inequivalence between the language versions, a DIF analysis was conducted. To identify linguistic differences between the language versions regarding voice, grammatical person, and sentence structure, task texts were analyzed. Finally, to investigate whether the differences in the linguistic features may be sources of inequivalence, correlations between the linguistic differences and DIF were analyzed.

In the analyses, no significant correlations between the differences in the linguistic features between the language versions of the tasks and the occurrence of DIF were found. Therefore, differences in the use of active and passive voice, grammatical person, and sentence structure are not likely to be sources of inequivalence between the English, German, and Swedish versions of these tasks.

Still, mathematics PISA tasks displaying DIF between the English, German, and Swedish versions were identified. The highest number of tasks displaying DIF was found when comparing the English and the Swedish versions where 32% of the tasks displayed DIF to different degree. About 29% of the tasks displayed DIF between the English and the German version and 17% of the tasks between the German and the Swedish version.

Since the statistical analyses in Study III revealed the occurrence of DIF but no correlations with the differences in linguistic features, other features of the tasks have to be sources of inequivalence. Not only linguistic differences between the language versions are possible sources of inequivalence, but also, for example, cultural or curricula differences between the language groups may cause inequivalence between the task versions. To find possible sources of inequivalence, in Study IV, students in Germany and Sweden worked with some of the tasks that were identified in Study III as displaying DIF between the German and the Swedish versions. The students were interviewed afterwards about their perceptions of obstacles and difficulties in the tasks.

Differences between the language versions in the use of uncommon words and complex sentences were identified as possible *linguistic* sources of inequivalence. A sentence that was formulated in a more difficult way in the Swedish text compared to the German version of the task was also found to possibly be a linguistic source of inequivalence. A possible non-linguistic source of inequivalence is the difference in how German and Swedish students dealt with information in the task text, not necessary to solve the task. The German students mentioned this information more often as disturbing or time consuming. This difference may occur due to *curricular* reasons, such as the type of tasks the students are used to work with. However, the results of study IV showed that *cultural* differences between the German and Swedish students do not seem to cause inequivalence in these tasks.

The combination of the results of the quantitative approach in Study III and the qualitative analysis in study IV in a mixed-methods design strengthen the finding that differences in the use of active and passive voice do not seem to be sources of inequivalence between the language versions of the tasks investigated. In Study III, no significant correlation of differences in the use of active and passive voice with the occurrence of DIF between the German and Swedish version of the tasks was found. This is in line with the qualitative analysis in Study IV that gave no evidence that students did have problems with passive voice in either language.

On the other hand, there were also results in Study III and IV that were not in line with each other. For example, differences in sentence structure between the language versions were not correlated to the occurrence of DIF in Study III, but in Study IV there were some differences between the student groups in how difficult they perceived some sentences to be. The reason for why students perceived sentences as difficult can vary, but some students commented also on the sentence structure. For example, one of the German students commented like this on a sentence: "It is a relatively complex sentence. It contains three commas and I wonder if it would be possible to put it in shorter, easier sentences."

4.3 Summary of results

In the following list, all results regarding different features of the tasks and their relations to unnecessary reading demands (measured by DRA), difficulties in solving, and inequivalence between language versions (measured by DIF) are summarized. The language versions are written in parentheses. S1 to S4 denote the study (I to IV, respectively) where the result was found.

- Word length
    - Negatively correlated to DRA (GER) S1
    - Not correlated to DRA (ENG, SWE) S1
    - Not correlated to difficulties in solving (ENG, GER, SWE) S1
- Uncommon words
    - Possible source of DRA (SWE) S2
    - Possible source of DIF (GER-SWE) S4
- Sentence length
    - Not correlated to DRA (ENG, GER, SWE) S1
    - Not correlated to difficulties in solving (ENG, GER, SWE) S1
- Sentence structure
    - Possible source of DRA (SWE) S2
    - Not correlated to DIF (ENG-GER, ENG-SWE, GER-SWE) S3
    - Possible source of DIF (GER-SWE) S4
- Task length
    - Not correlated to DRA (ENG, GER, SWE) S1

    - Not correlated to difficulties in solving (ENG, GER, SWE) S1
- Information density
    - Positively correlated to DRA (GER) S1
    - Not correlated to DRA (ENG, SWE) S1
    - Negatively correlated to difficulties in solving (GER, SWE) S1
    - Not correlated to difficulties in solving (ENG) S1
- Active/passive voice
    - Not correlated to DIF (ENG-GER, ENG-SWE, GER-SWE) S3
    - Possibly no source of DIF (GER-SWE) S4
- Second/third person
    - Not correlated to DIF (ENG-GER, ENG-SWE, GER-SWE) S3
- Curricular differences
    - Possible source of DIF (GER-SWE) S4
- Cultural differences
    - Possibly no source of DIF (GER-SWE) S4

5 Discussion

In this chapter, the results of the studies are discussed, mostly regarding findings coming from the combination and synthesis of the different studies. Some implications of the results for practice are given and desirable future research to further deepen the knowledge about the topics in this thesis is outlined.

5.1 Summary of main findings of the thesis

The findings of the studies in this thesis support the assumption that mathematics cannot be seen independent from language. In the studies included in this thesis, evidence was found that language can affect mathematics tasks in different ways. Firstly, linguistic features of the task texts can make the task more difficult to read and to solve, but which these features are, can differ between different natural languages. Secondly, when a task is given in different language versions, differences between the language versions can lead to inequivalence. That is, differences in difficulty can occur between the language versions that jeopardize the validity of results of multilanguage assessments. This inequivalence between the language versions can have different sources. In this thesis, curricular and linguistic differences were identified as possible sources of inequivalence. Still, also cultural differences between language groups may cause inequivalence, although this seems not to be the case with the German and Swedish versions of the six tasks investigated in Study IV in this thesis.

5.2 Linguistic features as sources of inequivalence and unnecessary reading demands

All four studies included in this thesis dealt in some way with the identification of linguistic features that can possibly make it harder for students to solve a mathematics task correctly. An example of an issue that can possibly make a task harder to read by increasing the task's unnecessary reading demand and that also may be a source of inequivalence between different language versions of a task, is the use of uncommon words. In Study II, the use of uncommon words showed to be an obstacle for students when reading a mathematics tasks and by that be a possible source for unnecessary reading demands in some mathematics tasks. This result is in line with findings in a quantitative study by Dyrvold et al. (2015), who found that measures of the occurrence of uncommon words in mathematics tasks are correlated with DRA. Differences in the use of uncommon words between language versions of tasks were also identified as possible sources of inequivalence between the different language versions of some tasks in Study IV. Differences in commonness of words between different language versions of tasks was also identified as sources of inequivalence in earlier studies (e.g., Roth et al., 2013).

Another linguistic feature that may be a source of both unnecessary reading demands and inequivalence is the use of complex sentences in the task text. In Study II, a complex sentence in a task showed to be an obstacle for students when reading the task. Also in Study IV, complex sentence structure was mentioned by some students as an obstacle and could be a possible source of inequivalence between the language versions. On the other hand, in Study III, differences in sentence structure between the language versions were not significantly correlated to DIF (as a measure of inequivalence). An explanation for the different results in Study III and IV regarding sentence structure could be that even if differences in sentence structure are not related to DIF generally, they may have some impact on inequivalence for some of the tasks. Another possible explanation could be that a sentence with equal formal complexity in both language versions (e.g., consisting of one main clause and two subordinate clauses) still can be easier to understand in one of the languages, for example, because of the mutual order of the clauses in the sentence (e.g., Allalouf, 2003). More research is needed to further investigate this question.

However, linguistic features that are related to DRA can reasonably be also sources of inequivalence. Two different scenarios are conceivable: If a linguistic feature is a source of unnecessary reading demands in both languages but appears only in one of the language versions of the task, the favor is for the version where it is absent. If the feature appears in both language versions of a task but is related to DRA in only one of the languages, this gives a favor to the language version where the feature is not a source of unnecessary reading demands. For example, the linguistic features examined in Study I (i.e., word length, sentence length, task length, and information density), were not related to DRA in tasks in English and Swedish, while some of them were related to DRA in German tasks. This could lead to a situation as described in the second scenario.

The fact that no correlation was found between linguistic features and DRA in Study I for the Swedish versions was somewhat unexpected. In an earlier study, Österholm and Bergqvist (2012b) had found correlations between DRA and word length and information density for mathematics tasks of PISA 2003 and 2006 in Swedish. This result was not reproduced for the tasks of PISA 2012 that were used in the current study. A possible reason is that these relations are not established for Swedish tasks in general. Also here, more research is desirable to further investigate this question.

5.3 Implications for practice

The results of the studies in this thesis have some implications for practice, for example when designing mathematics assessments, both in only one language

and in multilingual settings. Some of these implications are discussed in this section.

When designing tasks that are supposed to assess students' mathematical competence, it is necessary to avoid unnecessary reading demands in the tasks. Otherwise, there is a risk that the task to some degree assesses reading competence. To avoid unnecessary reading demands, one has to know which features of the task's text may cause unnecessary reading demands. The findings in this study show that simple advises as using short words or sentences, which are often used in readability formulas (e.g., Lenzner, 2014) and as indicators for high linguistic complexity of texts in general (e.g., Abedi et al., 2008), may not be enough to reduce unnecessary demands of reading ability in a task. Other linguistic features like, for example, familiarity of words or complexity of sentences have also to be considered (see Prediger et al., 2015). The same advice can be assumed to be valid also for other mathematics texts, like instructional texts in a textbook, examples, or exercises to avoid unnecessary obstacles for students working with these tasks. Thus, also in these types of texts, one have to be aware about the possibility of the occurrence of unnecessary reading demands.

Still, if a mathematics task is given in a real-world context, there is often some text (e.g., some background information) that is unnecessary to solve the task and that might be source of unnecessary reading demands. The positive effects of a real-world context of mathematics texts (see e.g., Boaler, 1993) and possible increased reading demands both have to be taken into consideration when designing tasks. Furthermore, mathematical literacy as defined by the OECD in the PISA framework includes the use of mathematics in "the context of a challenge or problem that arises in the real world" (OECD, 2012b, p. 25). That is, it is inevitable to put tasks in a real-world context to assess this kind of mathematical literacy, as it is done in the PISA tasks. The findings of study IV showed that students can react differently to this supplementary text. Some students just ignored it, others experienced it as unnecessary, time consuming, or disturbing. When designing mathematics tasks with problems set in a real-world context, it is necessary to find the right balance between the aim to asses this kind of mathematical literacy and possibly enhanced reading demands.

Another implication of the findings in this thesis is important to have in mind when designing multilanguage assessments and translating mathematics tasks into other languages. In this case, it is important to achieve a high degree of equivalence between the language versions. Still, equivalence in formal aspects as in the view of translation theory and equivalence in difficulty as in the view of test theory are not necessarily connected to each other, as shown by the following finding. The results of study III and IV both indicated that differences in the use of active and passive voice between the English, German, and Swedish language

versions of the mathematics PISA tasks probably are not sources of inequivalence. This is an interesting result, since passive voice often is said to increase reading difficulty and by that even difficulty of mathematics tasks. For example, Abedi, Lord, and Plummer (1997) found that changing from passive to active voice in mathematics tasks could lead to higher scores of students on these tasks. Also, the OECD recommends not to make changes in the use of active and passive voice when translating PISA-tasks to different languages (OECD, 2010). This recommendation can be questioned by the combined results of these studies if equivalence is seen in the view of test theory (Arffman, 2010), since the change of voice does not seem to have any impact on tasks' difficulty. Thus, a change between active and passive voice between language versions makes only a difference in the view of translation theory, which focuses on formal aspects of the texts.

5.4 More research to go further

The studies regarding reading demands (Study I & II) and the studies regarding equivalence (Study III & IV) showed some interesting paths where more research is needed to further shed light on the questions this thesis deals with.

As discussed in section 5.2, possible reasons for the inconsistency between the results of Study III and IV regarding the role of complex sentences in relation to DIF deserve further investigation. Also, reasons for the different results regarding linguistic features and their relation to DRA in Study I compared to earlier research (Österholm & Bergqvist, 2012b) should be investigated more.

To further investigate the role of linguistic features regarding DRA and difficulty of mathematics tasks, more qualitative analyses of Swedish tasks with high DRA should be conducted. For example, a comparison with tasks with low DRA would be helpful to find differences and similarities regarding linguistic features between the different tasks. Similar qualitative studies with tasks in English and German could help to find differences and similarities between tasks in different languages. A desirable next step is a triangulation with quantitative analyses of findings regarding linguistic features identified qualitatively as possible sources of unnecessary reading demands, that is, take a step in the QUAL-QUAN direction.

A further step would then be to investigate the relation of linguistic features and DRA in languages with a higher linguistic distance to the three languages examined in this study, which all are Germanic languages. Similar quantitative and qualitative studies like the ones in this thesis could help to find similarities and differences between the languages regarding this relation.

An interesting next step in the analysis of possible sources of inequivalence would be to include students from the USA in a qualitative study similar to Study IV, that is, analyzing task-based interviews. Since DIF was detected between the English (USA) version and both the German and Swedish versions of some PISA tasks, sources of inequivalence have to be identified to enable a higher degree of equivalence. For example, cultural differences may play a bigger role than between German and Swedish students. Also here, studies including more languages and students from different cultures could give interesting and important information about sources of inequivalence between different language versions of tasks. Furthermore, tasks without DIF between the language versions should be investigated to elicit whether there actually are less differences between the language versions with low DIF compared to tasks with higher level of DIF.

A final step would be to make changes in the language versions of the tasks to eliminate the issues identified as possible sources of inequivalence and then let students work with the reworded tasks. New DIF analyses could then show if DIF actually decreased, that is, whether the equivalence was enhanced.

5.5 Concluding remarks

This thesis makes a contribution to the understanding of the relation between language and mathematics by investigating how linguistic features of mathematics tasks can relate to difficulties in reading and solving the tasks. It also identifies some possible sources of inequivalence between different language versions of mathematics tasks by including "individuals who are representative of the examinee population for whom the assessment is designed" as advocated by Benítez, Padilla, Hidalgo Montesinos, and Sireci (2016). Combining quantitative and qualitative methods in a mixed methods design has showed to be a favorable way to identify features in mathematics tasks that can threaten the validity of assessments, both regarding if the task assesses what it is supposed to assess (mathematical ability, not reading ability) and if different language versions of a task are equivalent. More research is needed to further investigate these questions, and this thesis shows a promising way to go.

6 Acknowledgements/Tackord

Under min tid som doktorand på Institutionen för naturvetenskapernas och matematikens didaktik (NMD) på Umeå Universitet har jag hunnit lära mig så otroligt mycket och kunnat utvecklas mer än jag någonsin trodde. Allt detta hade inte varit möjligt utan alla fina människor runt omkring mig både på jobbet och i privatlivet. (Ja, även som doktorand har man lite privatliv kvar.)

Det hela började med att jag såg länken till en jobbannons på Carinas Facebooksida. Det söktes en doktorand inom pedagogiskt arbete med inriktning mot matematikdidaktik som förutom att vara allmän behörig även skulle ha goda kunskaper i tyska. Hur många sådana finns det här uppe i Norrland? – Tänkte jag. Jag sökte tjänsten och fick den till slut fast det höll på att strula till sig först, men då gick Johan L. in och såg till att det funkade. Tack för att ni fick mig att komma till NMD!

Sedan träffade jag mina handledare, Ewa och Magnus, och man (eller i alla fall jag) kan inte tänka mig bättre handledare till en vilsen doktorand. De båda hjälpte till att få struktur på arbetet och var alltid tillgängliga när det behövdes hjälp. Tack! Ni är bäst helt enkelt! När de gav mig den första artikeln som jag skulle läsa för att komma in i ämnet (Österholm & Bergqvist, 2012b) tänkte jag först – Jag fattar inget! Vad gör jag här? Lika bra att ge upp direkt! Men min nästa tanke var att jag är här för att lära mig. Ingen förväntar sig att jag redan kan allting från början. Och så mycket jag har lärt mig sedan dess tack vare er.
Bara någon vecka efter jag hade börjat såg jag för första gången någon försvara en avhandling. Det var Lotta som försvarade sin licentiatavhandling (Vingsle, 2014). Jag var väldigt imponerad och tänkte igen – Hur ska jag någonsin kunna lyckas med något sådant? Tre år senare hade jag förmånen att få Lotta som tredje handledare. Tack för all hjälp jag fick med ytterligare ett perspektiv på mitt arbete!

Hela min tid på NMD delade jag rum med Anders – enligt mig den perfekte rumskompisen. Vi gick många kurser tillsammans, hjälptes åt vid frågor och problem, snackade om oväsentligheter och tävlade i Seterra, men kunde också sitta tyst försjunkna i arbete en hel dag så att jag ibland inte ens visste om han var på rummet eller ej. (Han satt rakt mitt emot mig med bara två datorskärmar emellan.) Tack, Anders! Hoppas du klarar din sista tid på NMD utan mig.
Under mina första statistikkurs hade jag förutom Anders även Marléne som trogen medkämpe. Det var också hon som uppmuntrade mig att söka doktorandtjänsten från början fast hon inte var på samma institution. Vi kämpade inte bara på med statistiska samband, korrelationer, t-tester och chi-square utan utveck-

lade även Bland-Altman till Theens-Hedelin (Olafsdottir et al., 2016). Tillsammans med Jonas och Peter blev det också en hel del friskvårdtimmar i skidspår eller på löparbanan. Tack för blod(smak), svett(lukt) och (skratt)tårar!
Thank you, Candia for your input and support when reading my texts for the 50%-seminar. Och tack till Thomas för dina värdefulla tips och kommentarer som jag fick på mitt 90%-seminarium. Tack även alla andra doktorander och medarbetare på NMD! Tack för det underbart trevliga arbetsklimatet på institutionen, all hjälp och stöttning och alla skratt man har fått. Och tack för att ni har låtit mig tvinga er att åka skidor så mycket att vi vann UPIF:s skidserie fyra av fem år som jag har varit här. Bra jobbat!

Och Sara…! Vi har kämpat och slitit ihop den sista biten av vår doktorandtid. Vi hade tänkt gå i mål samtidigt, men nu blir jag klar strax före dig. Jag ser fram emot att få gå på din disputation snart. Och så får vi ju två fester istället för bara en! Tack för all stöttning, promenaderna, lunch- och fikasällskap, de oräkneliga skumtomtarna, helt enkelt: ”Tack för att du är en sådan fantastisk vän och kollega! Jag är glad att det var dig jag fick avsluta det här med.”

Tack alla elever och lärare i Sverige som har medverkat i min datainsamling. Und natürlich auch vielen Dank an die Schülerinnen und Schüler, Lehrerinnen und Lehrer an den Schulen in Deutschland, wo ich Interviews durchgeführt habe, sowie an das Ministerium für Bildung, Wissenschaft und Kultur des Landes Schleswig-Holstein, das mir die Untersuchung an den Schulen genehmigt hat.

Och som jag skrev ovan så behövs det också personer i privatlivet runt omkring en som på sitt sätt ser till att en doktorsavhandling kan bli klar. Zuallererst natürlich meine Eltern Karin und Dietmar Theens, die immer Vorbilder für mich waren. Meine Mutter ist mir den Weg als Lehrer vorausgegangen, und fast 50 Jahre nach meinem Vater (Theens, 1970) mache ich jetzt meinen Doktor. Und Danke auch an meine Schwestern Frugt und Solli, die ihren Teil mit dazu beigetragen haben, dass ich der bin, der ich bin. Danke auch an die besten Töchter der Welt, Evelina und Amalia! Jag tror hela NMD är tacksam för all gofika som räddades undan soptunnorna tack vare er och mina hungriga kollegor.

Ett stort tack till träningskompisarna i Stöcke TS Järnet. Det är viktigt att också hålla kroppen aktiv när man som doktorand mest sitter framför datorn eller läser artiklar. Tack vare er hann jag under doktorandtiden också med bl a 12 marathonlopp, 2 Ironman och en svensk klassiker. Tack för all träningspepp och energi!

Och sedan naturligtvis ett stort tack till Susanna för du stod ut med mig fast jag bara jobbade och tränade och inte hade tid för så mycket mer.

Umeå, i juni 2019

7 References

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377-398).

Abedi, J., Leon, S., Wolf, M. K., & Farnsworth, T. (2008). *Detecting test items differentially impacting the performance of ELL students*. Retrieved from

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance (CSE Report 429)*. Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

AERA. (2014). *Standards for educational and psychological testing*: American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational Psychological Testing.

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16*(1), 55-73.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the Causes of DIF in Translated Verbal Items. *Journal of educational measurement, 36*(3), 185-198.

Arffman, I. (2010). Equivalence of Translations in International Reading Literacy Studies. *Scandinavian Journal of Educational Research, 54*(1), 37-59.

Benítez, I., & Padilla, J.-L. (2014). Analysis of Nonequivalent Assessments across Different Linguistic Groups Using a Mixed Methods Approach: Understanding the Causes of Differential Item Functioning by Cognitive Interviewing. *Journal of Mixed Methods Research, 8*(1), 52-68. doi:10.1177/1558689813488245

Benítez, I., Padilla, J.-L., Hidalgo Montesinos, M. D., & Sireci, S. G. (2016). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education, 29*(1), 1-16.

Bergqvist, E., Theens, F., & Österholm, M. (2018). The role of linguistic features when reading and solving mathematics tasks in different languages. *The Journal of Mathematical Behavior, 51*, 41-55.

Boaler, J. (1993). The Role of Contexts in the Mathematics Classroom: Do they Make Mathematics More" Real"? *For the learning of mathematics, 13*(2), 12-17.

Bullock, J. O. (1994). Literacy in the language of mathematics. *The American Mathematical Monthly, 101*(8), 735-743.

Chen, F., & Chalhoub-Deville, M. (2016). Differential and long-term language impact on math. *Language Testing, 33*(4), 577-605.

Çikrikçi Demirtaşli, N., & Ulutaş, S. (2015). A Study on Detecting of Differential Item Functioning of PISA 2006 Science Literacy Items in Turkish and American Samples. *Eurasian Journal of Educational Research, 58*, 41-60.

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education*: Routledge.

Dorans, N. J., & Holland, P. W. (1992). DIF Detection and Description: Mantel-Haenszel and Standardization. *ETS Research Report Series, 1992*(1), i-40.

Dyrvold, A. (2016). The role of semiotic resources when reading and solving mathematics tasks. *Nordisk matematikkdidaktikk, 21*(3), 51-72.

Dyrvold, A., Bergqvist, E., & Österholm, M. (2015). Uncommon vocabulary in mathematical tasks in relation to demand of reading ability and solution frequency. *Nordisk matematikkdidaktikk, 20*(1).

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential Item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*(2), 24-35.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301-321.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA The MIT press.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis. *Scientific Studies of Reading, 5*(3), 239-256. doi:10.1207/S1532799XSSR0503_3

Goldin, G. A. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. In A. Kelly & R. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education* (pp. 517-545). Mahwah, NJ: Lawrence Erlbaum.

Grisay, A., De Jong, J., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement, 8*(3), 249-266.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

Huang, X., Wilson, M., & Wang, L. (2014). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology*(ahead-of-print), 1-13.

Johansson, B. G. (2004). *Matematikens historia*. Lund, Sweden: Studentlitteratur.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Johnson, T. P. (2006). Methods and frameworks for crosscultural measurement. *Medical care, 44*(11), S17-S20.

Kaiser, G., & Schwarz, I. (2003). Mathematische Literalität unter einer sprachlich-kulturellen Perspektive. *Zeitschrift für Erziehungswissenschaft, 6*(3), 357-377. doi:10.1007/s11618-003-0040-3

Koller, W. (2011). *Einführung in die Übersetzungswissenschaft* (8., neubearb. Aufl. ed.). Tübingen ; Basel: Francke.

Lenzner, T. (2014). Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty? *Sociological Methods & Research, 43*(4), 677-698.
Lithner, J., Bergqvist, E., Bergqvist, T., Boesen, J., Palm, T., & Palmberg, B. (2010). *Mathematical competencies: A research framework*. Paper presented at the The seventh mathematics education research seminar, Stockholm, January 26-27, 2010.
Michaelides, M. P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment Research & Evaluation, 13*(7).
Morgan, C., Craig, T., Schuette, M., & Wagner, D. (2014). Language and communication in mathematics education: an overview of research in the field. *ZDM, 46*(6), 843-853.
NCTM. (2000). *Principles and standards for school mathematics* (Vol. 1). Reston, VA: National Council of Teachers of Mathematics.
Niss, M., & Højgaard, T. (2011). *Competencies and mathematical learning - Ideas and inspiration for the development of mathematics teaching and learning in Denmark* (Vol. 485). Roskilde, Denmark: Roskilde University.
NRC. (2001). *Adding it up: Helping children learn mathematics*: National Academies Press.
O'Grady, K. F., Karen; Servage, Laura; Khan, Gulam. (2018). *PCAP 2016 Report on the Pan-Canadian Assessment of Reading, Mathematics, and Science*. Retrieved from Toronto, ON: https://www.cmec.ca/Publications/Lists/Publications/Attachments/381/PCAP-2016-Public-Report-EN.pdf
OECD. (2010). Translation and Adaption Guidelines for PISA 2012. Budapest, Hungary: OECD.
OECD. (2012a). Data base - PISA 2012. Retrieved from http://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm
OECD. (2012b). PISA 2012 Assessment and Analytical framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy: OECD Publishing.
OECD. (2014). PISA 2012 Results in Focus: What 15-year-olds know and what they can do with what they know: OECD Publishing.
Olafsdottir, A. S., Hörnell, A., Hedelin, M., Waling, M., Gunnarsdottir, I., & Olsson, C. (2016). Development and validation of a photographic method to use for dietary assessment in school settings. *PloS one, 11*(10), e0163970.
Panou, D. (2013). Equivalence in translation theories: A critical evaluation. *Theory and Practice in Language Studies, 3*(1), 1-7.
Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child development, 78*(4), 1255-1264.
Prediger, S., Wilhelm, N., Büchter, A., Gürsoy, E., & Benholz, C. (2015). Sprachkompetenz und Mathematikleistung – Empirische Untersuchung sprachlich bedingter Hürden in den Zentralen Prüfungen 10. *Journal für Mathematik-Didaktik, 36*(1), 77-104. doi:10.1007/s13138-015-0074-0
PRIM-gruppen. (2019). Tidigare ämnesprov för årskurs 9. Retrieved from https://www.su.se/primgruppen/matematik/%C3%A5rskurs-9/tidigare-prov

Riccomini, P. J., Smith, G. W., Hughes, E. M., & Fries, K. M. (2015). The language of mathematics: The importance of teaching and learning mathematical vocabulary. *Reading & Writing Quarterly, 31*(3), 235-252.

Roe, A., & Taube, K. (2006). How Can Reading Abilities Explain Differences in Maths Performances? In J. Mejding & A. Roe (Eds.), *Northern Lights on PISA 2003: A Reflection from the Nordic Countries* (pp. 129-141). Copenhagen: Nordic Council of Ministers.

Roth, W.-M., Oliveri, M. E., Sandilands, D. D., Lyons-Thomas, J., & Ercikan, K. (2013). Investigating Linguistic Sources of Differential Item Functioning Using Expert Think-Aloud Protocols in Science Achievement Tests. *International Journal of Science Education, 35*(4), 546-576.

Schleppegrell, M. J. (2007). The Linguistic Challenges of Mathematics Teaching and Learning: A Research Review. *Reading & Writing Quarterly, 23*(2), 139-159. doi:DOI: 10.1080/10573560601158461

Shanahan, T., & Shanahan, C. (2008). Teaching Disciplinary Literacy to Adolescents: Rethinking Content-Area Literacy. *Harvard Educational Review, 78*(1), 40-59.

Theens, D. (1970). *Die Entwicklung der Gas- und Elektrizitätsversorgung Schleswig-Holstein/Hamburgs.* (Dr rer nat), Christian-Albrechts-Universität zu Kiel, Kiel, Germany.

Wakefield, D. V. (2000). Math as a second language. *The Educational Forum, 64*(3), 272-279.

Vetenskapsrådet. (2011). *God forskningssed*: Vetenskapsrådet.

Vingsle, C. (2014). *Formative assessment: Teacher knowledge and skills to make it happen.* Umeå universitet.

Yildirim, H. H., & Berberoğlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing, 9*(2), 108-121.

Österholm, M., & Bergqvist, E. (2012a). Methodological issues when studying the relationship between reading and solving mathematical tasks. *Nordic Studies in Mathematics Education, 17*(1), 5-30.

Österholm, M., & Bergqvist, E. (2012b). *What mathematical task properties can cause an unnecessary demand of reading ability?* Paper presented at the Proceedings of Norma 11, The Sixth Nordic Conference on Mathematics Education in Reykjavík, May 11-14, 2011.