Funktionales Denken und die Rolle des Geschlechts: Explorative Analyse quantitativer Testdaten

MARCEL KLINGER, ESSEN

***Zusammenfassung**: Im Rahmen der FALKE-Erhebung zum Funktionalen Denken wurde ein Leistungstest zu funktionalen Zusammenhängen entwickelt und in Nordrhein-Westfalen mit über 3000 Schülerinnen und Schülern durchgeführt. Gerade für die Sekundarstufe werden im Rahmen von Meta-Studien mathematischen Leistungstests häufig geschlechtsspezifische Effekte zu Gunsten des männlichen Geschlechts attestiert. Der vorliegende Beitrag analysiert solche Effekte für die Stichprobe des FALKE-Tests und exploriert, welche Items besonders betroffen sind. Es lassen sich Merkmale von Aufgaben im Bereich des Funktionalen Denkens identifizieren, die besonders geschlechtssensitiv wirken.*

***Abstract:** Within the FALKE-study an achievement test for functional thinking has been administrated to more than 3000 students in the German federal state of North Rhine-Westphalia. Especially for the secondary school level meta studies find significant gender specific effects in favor of male test attendants. This article analyzes such effects for the sample of the FALKE-test and explores which items are affected in particular. Item characteristics which favor gender sensitivity of tasks in the field of functional thinking are identified.*

1. Einleitung

Der im Rahmen der Dissertation von Klinger (2018) entstandene FALKE-Test fokussiert das inhaltliche Verständnis von Schülerinnen und Schülern zu Beginn der Oberstufe in den Bereichen Funktionenlehre und frühe Analysis.[1] Das Testinstrument wurde im Rahmen der entsprechenden Qualifikationsarbeit entwickelt und validiert. Hierbei kam eine Stichprobe von über 3000 Schülerinnen und Schülern des ersten Oberstufenjahres – der sog. Einführungsphase – in Nordrhein-Westfalen zum Einsatz.[2]

Wie für die meisten mathematischen Leistungstests stellt sich auch für diese Stichprobe ein signifikanter Leistungsvorsprung zu Gunsten des männlichen Geschlechts ein. Die FALKE-Erhebung ordnet sich somit in eine Reihe mathematischer Leistungsstudien ein, bei denen Mädchen bzw. Frauen im Schnitt signifikant schlechter abschneiden als entsprechende männliche Probanden. I. d. R. wird bei der Analyse dieser sog. *Gender Gap* vor allem Gesamtleistung als geschlechtsspezifisch-variierende Variable betrachtet. Mitunter schwankt ein entsprechender Effekt aber deutlich zwischen einzelnen Items, so auch im Rahmen des FALKE-Tests. Zwar ergeben sich auch hier bei Betrachtung der Lösungsquoten für nahezu alle Einzelitems entsprechende Vorteile zu Gunsten der Jungen, jedoch reichen diese von vernachlässigbar kleinen Effekten bis hin zu stärkeren Effekten. Letztere äußern sich etwa in einer Differenz von 19.7 Prozentpunkten zwischen den Lösungsquoten der männlichen und weiblichen Substichprobe einer spezifischen Leistungstestaufgabe. Das entsprechende Item „Kegelfüllung“ ist in Abbildung 1 dargestellt.

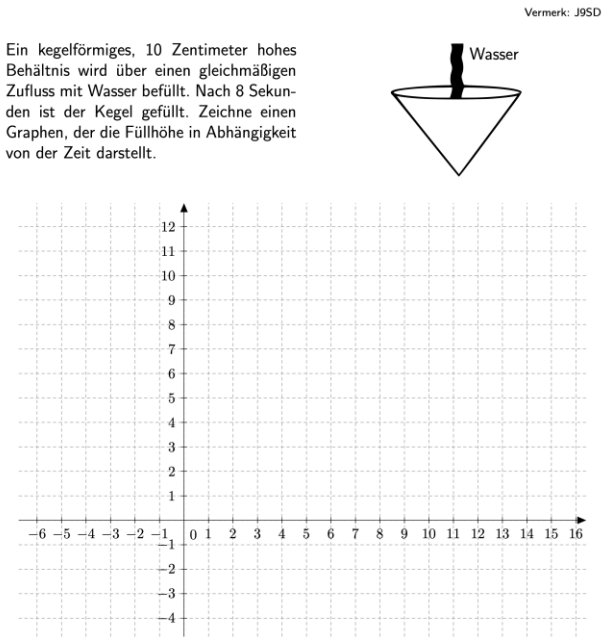

Abb. 1: Aufgabe „Kegelfüllung“ (Kennung J9SD) des FALKE-Tests (Klinger, 2018)

Konkret musste zur Lösung des entsprechenden Items ein konkaver Füllgraph qualitativ korrekt skizziert werden. Dieser musste zudem den Punkt (8|10) beinhalten und dort enden bzw. ab dort konstant fortgeführt werden. Hierbei erzielte die männliche Substichprobe 752 korrekte Lösungen bei 1584 Probanden (47.5 %); die weibliche 445 korrekte Lösungen bei 1602 Probandinnen (27.8 %). Es zeigt sich also ein nicht unerheblicher Effekt bereits auf Ebene einzelner Items, der aus einzelnen Aspekten der Aufgabengestaltung aber auch in den zur Bearbeitung notwendigen Teilkompetenzen rühren kann.

Der vorliegende Artikel versucht am Beispiel des FALKE-Tests solche Teilaspekte von Aufgaben aus dem Bereich Funktionenlehre und frühe Analysis herauszuarbeiten, die von besonderer geschlechtsspezifischer Relevanz zu sein scheinen (vgl. Klinger & Barzel, 2018a). Es steht somit nicht zuletzt die Rolle des Geschlechts beim Funktionalen Denken im Fokus.

Im weiteren Verlauf wird zunächst allgemein zum Forschungsstand bezüglich des Zusammenhangs von Mathematikleistung und Geschlecht berichtet. Dort, wo spezifische Forschungserkenntnisse zu einschlägigen mathematischen Inhalten vorliegen, werden auch diese skizziert. Im zweiten Teil des Beitrags werden sodann entsprechende Effekte anhand der FALKE-Erhebung untersucht. Hierbei steht die Frage nach der Bedeutung solcher Effekte für den Bereich des Funktionalen Denkens im Mittelpunkt. Es werden weiterhin Aspekte auf Ebene einzelner Aufgaben herausgearbeitet, welche geschlechtsspezifische Effekte besonders begünstigen.

2. Mathematikleistung und Geschlecht

Es lässt sich eine Vielzahl an Forschungsarbeiten finden, die die Wirkung des Einflussfaktors Geschlecht auf die Ergebnisse mathematischer Leistungstests untersuchen. Im weiteren Sinne ist damit jede Art schriftlicher mathematischer Leistungserhebung gemeint; im engeren Sinne „Tests, bei denen Aufgaben bearbeitet werden müssen, deren Bearbeitung als richtig oder falsch gewertet wird“ (Eid & Schmidt 2014, S. 417).

Dabei zeichnet sich in mathematischen Leistungstests meist ein geringfügiger Effekt zu Gunsten des männlichen Geschlechts innerhalb nationaler wie internationaler Leistungsstudien ab. Der entsprechende Sachverhalt ist innerhalb einschlägiger Literatur spätestens seit den 1960er Jahren bekannt (Fennema, 1974; Fennema & Sherman, 1978). Für Klieme gehört dieser Leistungsvorsprung, den männliche gegenüber weiblichen Probanden genießen, sogar zu den „am deutlichsten ausgeprägten und am besten dokumentierten Befunden über Geschlechtsunterschiede im Bereich der Psychologie“ (Klieme, 1986, S. 133).

Der vorliegende Artikel fokussiert vor allem geschlechtsspezifische Effekte im Themenbereich Funktionen und Analysis und somit beim Funktionalen Denken. Dennoch soll zunächst die Bedeutung entsprechend geschlechtssensitiv ausfallender Leistungserhebungen im Allgemeinen geklärt werden.

2.1 Effektstärke geschlechtsspezifischer Differenzen

Um entsprechende Stärken eines etwaig auftretenden geschlechtsspezifischen Leistungseffekts auch über unterschiedliche Studien und Tests hinweg vergleichen zu können, wird i. d. R. das Effektstärkenmaß d nach Cohen (1988, S. 20 ff.) verwendet. Der Koeffizient wird dabei berechnet, indem der Mittelwert (der Testleistung) der weiblichen Probanden von jenem der männlichen Probanden subtrahiert wird. Diese Differenz wird sodann durch die gemeinsame (gepoolte) Standardabweichung geteilt. Mit anderen Worten stellt d die geschlechtsspezifische Differenz in Vielfachen der Standardabweichung dar (vgl. Hyde, 2005, S. 582).

Hierbei implizieren positive Werte von d dann einen gemessenen Leistungsvorsprung der Jungen bzw. Männer, negative Werte einen Leistungsvorsprung der Mädchen bzw. Frauen. Hierbei bewertet Hyde (2005) Werte von d, die betragsmäßig zwischen 0.00 und 0.10 liegen als „close-to-zero“. Werte zwischen 0.10 und 0.35 gelten als klein, Werte zwischen 0.35 und 0.65 als mittel, Werte zwischen 0.65 und 1.00 als groß sowie Werte über 1.00 als sehr groß (vgl. Brunner et al., 2011).

Im Rahmen mathematischer Leistungstests fallen Effektstärken i. d. R. klein bzw. nahe bei null, jedoch meist zu Gunsten der männlichen Probanden aus. Eine entsprechende Zusammenfassung einiger Meta-Studien, entsprechender Ergebnisse der großen internationalen TIMSS- und PISA-Untersuchungen sowie relevante Ergebnisse ausgewählter innerdeutscher Leistungsstudien sind in Tabelle 1 dargestellt. Die Tabelle geht in ihrer ursprünglichen Form auf eine Veröffentlichung von Brunner et al. (2011) zurück und wurde u. a. um die neuesten zur Verfügung stehenden Daten der TIMS- und PISA-Studie sowie um Studien aus dem deutschsprachigen Raum (namentlich der IQB-Ländervergleich 2012 sowie die Hamburger KESS-Studie) ergänzt. Da in der vorliegenden Abhandlung vor allem der Übergang zwischen den Sekundarstufen fokussiert wird, sollen auch vornehmlich Effektstärken innerhalb der entsprechenden Jahrgänge wiedergegeben werden.

Die Tabelle beginnt mit einer Meta-Analyse von Hyde, Fennema und Lamon (1990), welche insgesamt 100 Einzelstudien zusammenfasst und somit die umfangreichste quantitative Überblicksarbeit bis zum Jahr 1990 darstellt (vgl. Brunner et al., 2011, S. 181). Hieraus resultiert eine mittlere Effektstärke von $d = -0.05$, womit sich insgesamt wohl erstmals ein Leistungsvorsprung für das weibliche Geschlecht zeigt (vgl. Köller & Klieme, 2000, S. 373 f.). Betrachtet man hingegen lediglich Studien, die sich auf Schülerinnen und Schüler der High School im Alter

von 15 bis 18 Jahren beziehen, ergibt sich auch hier ein Effekt zu Gunsten des männlichen Geschlechts (dargestellt in der Tabelle). Eine Folgeuntersuchung von Lindberg et al. (2010) fokussiert den Zeitraum von 1990 und 2007 und betrachtet insgesamt 242 Einzelstudien mit fast 1,3 Millionen Probanden.

Auch in dieser Studie zeigt sich mit Blick auf Erhebungen in der Sekundarstufe ein Leistungsvorteil für Jungen von $d = 0.23$ anhand von 110 Einzelstudien, während sich bei globaler Betrachtung lediglich eine Effektstärke von $d = 0.05$ einstellt. Beide Meta-Analysen fassen dabei vornehmlich angloamerikanische Arbeiten zusammen.

Eine aktuellere Meta-Analyse von Reilly et al. (2015) umfasst mit einer Zeitspanne von 1990 bis 2011 zwar einen ähnlichen Zeitraum, fokussiert aber vor allem das US-amerikanische *National Assessment of Educational Progress* (NAEP). Hier zeigt sich eine gemittelte Effektstärke von $d = 0.10$ innerhalb des zwölften Jahrgangs, die auf einen Datensatz von

Studie	d	Alter
Meta-Analysen (überwiegend anglo-amerikanischer Raum)		
Hyde et al. (1990): 53 Einzelstudien, Zeitraum vor 1990	0.29	15–18 J.
Lindberg et al. (2010): 110 Einzelstudien, Zeitraum 1990 bis 2007	0.23	14–18 J.
Reilly et al. (2015): NAEP-Datensatz, Zeitraum 1990 bis 2011	0.10	12. Jg.
TIMSS (Trends in the International Mathematics and Science Study)		
1995: gemittelt über alle Teilnehmerstaaten	0.08	8. Jg.
1995: nur Deutschland	0.03	8. Jg.
1999: gemittelt über alle Teilnehmerstaaten	0.04	8. Jg.
2003: gemittelt über alle Teilnehmerstaaten	0.01	8. Jg.
2007: gemittelt über alle Teilnehmerstaaten	0.00	8. Jg.
2011: gemittelt über alle Teilnehmerstaaten	0.04	8. Jg.
TIMSS-Erhebung zur mathematischen und naturwissenschaftlichen Bildung am Ende der Schullaufbahn (Köller & Klieme, 2000, S. 402)		
1995: nur Deutschland (Skala „Voruniversitäre Mathematik“)	0.33	13. Jg.
1995: davon in Grundkursen	0.06	13. Jg.
1995: davon in Leistungskursen	0.29	13. Jg.
PISA (Programme for International Student Assessment)		
2000: gemittelt über alle Teilnehmerstaaten	0.11	15 J.
2000: nur Deutschland	0.15	15 J.
2003: gemittelt über alle Teilnehmerstaaten	0.11	15 J.
2003: nur Deutschland	0.09	15 J.
2006: gemittelt über alle Teilnehmerstaaten	0.11	15 J.
2006: nur Deutschland	0.20	15 J.
2009: gemittelt über alle Teilnehmerstaaten	0.12	15 J.
2009: nur Deutschland	0.16	15 J.
2012: gemittelt über alle Teilnehmerstaaten	0.12	15 J.
2012: nur Deutschland	0.15	15 J.
IQB-Ländervergleich 2012 (Schroeders et al., 2013, S. 259 ff.)		
2012: Gymnasien	0.28	9. Jg.
2012: andere Schulen	0.24	9. Jg.
2012: gemittelt über alle Schulformen	0.16	9. Jg.
2012: gemittelt über alle Schulformen in Nordrhein-Westfalen	0.32	9. Jg.
KESS 10/11 (Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der Sekundarstufe I und zu Beginn der gymnasialen Oberstufe) (Ivanov, 2011, S. 86)		
2009: Gymnasien	0.45	10./11. Jg.
2009: Gesamtschulen	0.33	10./11. Jg.
2009: Realschulen	0.36	10./11. Jg.
2009: gemittelt über alle Schulformen	0.29	10./11. Jg.

Tab. 1: Geschlechtsspezifische Effekte in mathematischen Leistungsstudien bei Schülerinnen und Schülern in der Sekundarstufe (abgewandelt übernommen und erweitert nach Brunner et al., 2011; TIMSS, 2011 ergänzt nach Mullis et al., 2012, S. 70; PISA, 2012 ergänzt nach OECD, 2014, S. 329)

104900 Schülerinnen und Schülern zurückgeht. Für die Jahrgänge 4 und 8 berichten die Autoren hingegen kleinere Effektstärken von $d = 0.07$ respektive $d = 0.04$ (vgl. Reilly et al., 2015, S. 651).

Weiterhin werden Daten aus den TIMS-Studien dargestellt. Da sich Deutschland im Bereich der Sekundarstufe seit 1995 nicht mehr an dem entsprechenden Programm für diese Jahrgangsstufe beteiligt (vgl. Brunner et al,. 2011, S. 182), werden ab 1999 ausschließlich internationale Werte als Mittelung über alle Teilnehmerstaaten berichtet. Insgesamt zeigen sich im Rahmen der TIMS-Studie deutlich geringere Effektstärken. Das letzte deutsche Datum für den achten Jahrgang aus dem Jahr 1995 kann mit einer Effektgröße von $d = 0.03$ als vernachlässigbar betrachtet werden. Die Effektstärke für die im Rahmen der TIMSS-Untersuchung durchgeführten Erhebung am Ende der gymnasialen Oberstufe im 13. Jahrgang fällt hingegen deutlich höher aus. Hier geben Köller und Klieme (2000, S. 381) eine Effektstärke von $d = 0.33$ auf der Skala „Voruniversitäre Mathematik“ für alle Schulformen mit gymnasialer Oberstufe an. Zu beachten ist hierbei jedoch, dass diese deutlich unterschiedlich ausfällt, je nachdem welche Kursform betrachtet wird. So zeigt sich in Leistungskursen ein Effekt von $d = 0.29$, während die geschlechtsspezifischen Differenzen in Grundkursen mit $d = 0.06$ im nicht-signifikanten Bereich liegen. Für die Autoren bieten sich diesbezüglich zwei unterschiedliche Interpretationen an: Einerseits besteht die Möglichkeit, dass ein nicht-trivialer Anteil an Frauen zwar das Potenzial für die Wahl eines Leistungskurses gehabt hätte, die betreffenden Personen aber z. B. aufgrund fehlenden Interesses oder vorherrschender Geschlechterstereotype eine Entscheidung zugunsten eines Grundkurses fällten. Andererseits könne es sich bei den männlichen Studienteilnehmern im Grundkurs auch um eine Negativ-Auslese handeln (vgl. Köller & Klieme, 2000, S. 403).

Ebenfalls in der Tendenz recht gering, jedoch intensiver als für den in TIMSS untersuchten deutschen Oberstufenjahrgang, zeigen sich geschlechtsspezifische Effekte innerhalb der PISA-Untersuchungen. Hierbei ergeben sich i. d. R. kleine Effektstärken zwischen $d = 0.10$ und $d = 0.20$. Tendenziell fallen geschlechtsspezifische Effekte zudem für die deutschen Fünfzehnjährigen (mit Ausnahme des Erhebungsjahres 2003) etwas deutlicher als im internationalen Mittel aus.

Eine reininnerdeutsche Vergleichsstudie ist in Form des IQB-Ländervergleichs von 2012 in der Tabelle aufgeführt. Diese wurde vor allem als nationales Monitoring-System zur Umsetzung der Bildungsstandards, welche im Rahmen des „PISA-Schocks“ um die Jahrtausendwende etabliert wurden, eingeführt (vgl. Pant et al., 2013). Betrachtet man die untersuchte Stichprobe in ihrer Gesamtheit, stellt sich ein kleiner Effekt von $d = 0.16$ zu Gunsten der Jungen ein, was einem Lernvorsprung der Jungen von etwa zwei Dritteln eines Schuljahres entspricht (vgl. Schroeders et al., 2013, S. 258). Zerlegt man die Stichprobe in die Gruppe der Gymnasiasten sowie die Gruppe der sonstigen Schulformen stellen sich jeweils höhere Effektstärken von $d = 0.28$ bzw. $d = 0.24$ ein. Auf diesen scheinbaren Widerspruch wird noch einmal in Abschnitt 2.2 eingegangen (sog. *Simpson-Paradoxon*).

Der Ergebnisbericht zur IQB-Studie enthält zudem auch Daten auf Ebene der einzelnen Länder. Hier zeigt sich, dass sich für das Land Nordrhein-Westfalen (in welchem auch der FALKE-Test durchgeführt wurde) mit $d = 0.32$ der größte geschlechtsspezifische Effekt einstellt. Insgesamt zeigen aber alle Länder geschlechtsspezifische Auffälligkeiten zu Ungunsten der Mädchen. Eine Ausnahme bildet Hessen, wo sich ein nicht-signifikanter Effekt zu Ungunsten der Jungen abzeichnet.

Mit dem Projekt „Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der Sekundarstufe I und zu Beginn der gymnasialen Oberstufe“ (KESS 10/11) enthält die Tabelle eine weitere innerdeutsche Studie, in der die Freie und Hansestadt Hamburg einen Jahrgang längsschnittlich von der Grundschule bis zum etwaigen Abitur untersucht. Die dargestellten Daten aus dem Jahr 2009 fokussieren die Schülerinnen und Schüler beim Übergang in die Oberstufe, so dass die vergleichsweise kleine Studie im Rahmen dieser Abhandlung von besonderer Bedeutung ist. Hier zeigt sich gerade am Gymnasium eine vergleichsweise hohe Effektstärke von $d = 0.45$. An den anderen betrachteten Schulformen fällt der entsprechende Wert hingegen etwas geringer aus. Am geringsten fällt die Effektstärke auch hier aus, betrachtet man alle Schulformen gemeinsam. Diese liegt dann bei $d = 0.29$.

Insgesamt zeigen sich deutlich häufiger geschlechtsspezifische Effekte zu Gunsten des männlichen Geschlechts. Die dargestellten Daten lassen vermuten, dass der Effekt gegen Ende der Sekundarstufe I und in der Oberstufe besonders ausgeprägt ist. Die Meta-Studie von Hyde et al. (1990) lässt zudem erkennen, dass mit zunehmendem Alter der Probanden sich auch geschlechtsspezifische Differenzen der Mathematikleistung vergrößern. So errechnen die Forscherinnen in der Altersgruppe der Fünf- bis Zehnjährigen eine mittlere Effektstärke von $d = -0.06$ und für die Elf- bis Vierzehnjährigen von $d = -0.07$. Diese für Mädchen günstige Ausgangslage verschiebt sich in der Altersgruppe der Fünfzehn- bis Achtzehnjährigen auf $d = 0.29$ zu Gunsten der Jungen bzw. Männer

und nimmt für die Altersgruppe der Neunzehn- bis Fünfundzwanzigjährigen noch weiter zu (d = 0.41), bis sie schließlich in der Gruppe der Übersechsundzwanzigjährigen mit d = 0.59 ihren Höhepunkt erreicht (vgl. Hyde et al., 1990, S. 148).

Dass geschlechtsspezifische Differenzen in mathematischen Leistungsstudien entlang der einzelnen institutionellen Bildungsetappen zunehmen, zeigen auch zahlreiche andere Autorinnen und Autoren (z. B. Contini et al., 2017; Robinson & Lubienski, 2011; Fryer & Levitt, 2010; Penner & Paret, 2008). So weisen insbesondere Studien in der Primarstufe eher geringe Effektstärken auf, während im Bereich der Sekundarstufe oder gar der universitären Bildung deutlich höhere Werte zu verzeichnen sind. Ebenfalls scheint auch die Art der Bildungseinrichtung sowie die jeweils betrachtete Kursform eine Rolle zu spielen. So fallen entsprechende Effektstärken für Studien aus dem deutschen Raum an Gymnasien höher aus als an sonstigen Schulformen. Zudem zeigen Schülerinnen und Schüler in Leistungskursen größere Differenzen als in Grundkursen.

2.2 Zur Variation der aufgezeigten Effektstärken und methodische Kritik

Messen einzelne Items eines Tests nicht nur die fokussierten Eigenschaften, also etwa mathematische Leistung, sondern zu einem gewissen Grad auch andere personenbezogene Eigenschaften (etwa Muttersprache, ethnische Zugehörigkeit oder wie in der hier betrachteten Situation das Geschlecht), deren Erhebung aber eigentlich nicht intendiert ist, so wird testtheoretisch von *Differential Item Functioning* (DIF) gesprochen (vgl. Lord, 1980, S. 212 ff.). Entsprechend definieren Embretson und Reise (2009) DIF wie folgt:

> DIF is said to occur when a test item does not have the same relationship to a latent variable (or multidimensional latent vector) across two or more examinee groups. (Embretson & Reise, 2009, S. 251)

Treten solche Effekte verstärkt auf, kann das Ergebnis eines gesamten Tests verzerrt werden, so dass auch von *Differential Test Functioning* (DTF) gesprochen wird (vgl. Linacre, 2002; Raju et al., 1995).

Innerhalb einschlägiger Literatur wird in aller Regel deutlich, dass DIF oder auch DTF einzig als qualitative Unzulänglichkeit des Testinstruments betrachtet wird. Gerade in großen Leistungsstudien wie PISA und TIMSS werden solche Items, die sich als besonders geschlechtssensitiv herausstellen, im Rahmen von vorgeschalteten Pilotierungsstudien aus dem Aufgabenpool entfernt (z. B. Adams & Carstensen, 2002, S. 154). Entsprechend stellen die in solchen Studien beobachteten geschlechtsspezifischen Differenzen wohl eher eine Unter- als eine Obergrenze hinsichtlich ihrer Effektstärken dar (vgl. Brunner et al., S. 199). Beim Ausschluss von Items aufgrund vermeintlicher DIF-Effekte muss also sorgfältig abgewogen werden zwischen einem Fehlverhalten des zu konzipierenden Testinstruments und tatsächlicher innerhalb der untersuchten Population vorhandener Merkmale, die korrekt erfasst werden. Letztlich bietet das Stattfinden oder Nicht-Stattfinden solcher Pilotierungen und entsprechender vorweggeschalteter DIF-Analysen einen methodischen Erklärungsansatz für die z. T. deutlich schwankenden Effektstärken der in Abschnitt 2.1 dargestellten Studien.

Darüber hinaus sind die von einzelnen Studien erhobenen Leistungswerte nicht ohne Weiteres gleichzusetzen. So erheben bereits PISA und TIMSS ihres Forschungsschwerpunktes entsprechend unterschiedliche, wenngleich sicherlich verwandte, mathematische Leistungsdimensionen. Während PISA mit dem Konzept der „mathematical literacy“ vor allem mathematische Kompetenzen in Alltagssituationen fokussiert, erhebt TIMSS „mathematisches Wissen“ und orientiert sich dabei deutlich stärker an nationalen Curricula (vgl. Leder & Forgasz, 2018, S. 691).

Nicht zuletzt hängt die Stärke erhobener geschlechtsspezifischer Effekte auch vom verwendeten statistischen Modell ab, mit denen die gemessene Leistung erfasst wird. So konnten etwa Brunner et al. (2011) für ein Nested-Faktormodell, in welchem die erhobene Leistung nicht nur von mathematischer Kompetenz, sondern auch der allgemeinen Intelligenz der Probanden statistisch abhängig ist, zeigen, dass sich so deutliche größere geschlechtsspezifische Effekte einstellen.

Letztlich ist auch die Stichprobenbildung bzw. die Auswahl einzelner Subgruppen dieser von besonderer Bedeutung für die Stärke (oder sogar das generelle Auftreten) entsprechender geschlechtsspezifischer Effekte. So lassen sich regelmäßig (Meta-)Studien finden, die zu dem Schluss kommen, dass kein geschlechtsspezifischer Effekt für mathematische Leistungstests existiert (z. B. Hyde et al., 2008). Mitunter ermitteln einzelne Studien auch einen gegenteiligen Effekt (z. B. Tartre & Fennema, 1995). Für Büchter (2010) liegt dies aber i. d. R. „an der Stichprobenziehung, an den getesteten Bereichen oder an strukturell verzerrenden Rahmenbedingungen“ (Büchter, 2010, S. 48). Als ein Beispiel strukturell verzerrender Rahmenbedingungen führt er eine der TIMSS-Erhebungen an. Hier heißt es:

> Mädchen erreichen in Mathematik und Physik in allen Schulformen schwächere Leistungen als Jungen. [...] Bei der Betrachtung der Leistungsbilanz von Jungen und Mädchen auf der Ebene des gesamten Altersjahrgangs treten im Fach Mathematik keine und im Fach

> Physik kleinere Leistungsunterschiede zwischen den Geschlechtern auf als in den einzelnen Schulformen. Dies ist ausschließlich eine Folge der höheren gymnasialen Bildungsbeteiligung von Mädchen [...]. (Baumert & Lehmann, 1997, S. 26)

Dieses scheinbar widersprüchliche Bild stellt ein sog. *Simpson-Paradoxon* dar (nach Simpson, 1951): Teilt man eine Gesamtstichprobe in unterschiedliche Gruppen auf (z. B. nach Schulform), ist es möglich, dass sich innerhalb jeder Gruppe dasselbe Bild ergibt, das Gesamtbild über alle Gruppen hinweg jedoch davon abweicht. Entsprechend zeigt sich innerhalb jeder Schulform hier zwar ein signifikanter geschlechtsspezifischer Unterschied, nicht jedoch über alle Schulformen hinweg. Dies kommt durch die gegenüber Jungen höhere Quote von Mädchen an Gymnasien und den Umstand, dass Gymnasiasten i. A. bessere Leistungen als Schülerinnen und Schüler entsprechenden Alters anderer Schulformen erzielen, zustande (vgl. Büchter, 2004; Dorans & Holland, 1993). So kommt bei den in Abschnitt 2.1 dargestellten Daten ein solches Simpson-Paradoxon neben der bereits angesprochenen IQB-Erhebung etwa bei der KESS-Erhebung als Erklärungsmodell in Frage. Betrachtet man hier nicht nur das Gymnasium, sondern alle Schulformen gemeinsam, sinkt die Effektstärke von $d = 0.45$ auf $d = 0.29$, obwohl sie an Gesamt- und Realschulen zwar geringer als am Gymnasium, aber dennoch oberhalb von $d = 0.29$ ausfällt. Da Bildungssysteme typischerweise in verschiedene Schulformen, etc. aufgeteilt sind, sollte bei der Analyse von Leistungsdaten ein mögliches Simpson-Paradoxon daher also stets mitgedacht werden (vgl. Büchter, 2010, S. 48).

2.3 Abhängigkeiten vom mathematischen Teilgebiet und mögliche Kovariaten entsprechender Effekte

In Abschnitt 2.1 wurde deutlich, dass die Stärke geschlechtsspezifischer Effekte in mathematischen Leistungsstudien einerseits vom betrachteten Jahrgang, aber auch von Schulform und Kursform abhängen. Neben diesen eher institutionellen Rahmenbedingungen lassen sich in der Literatur weitere Aspekte identifizieren, die die Stärke geschlechtsspezifischer Effekte in mathematischen Leistungsstudien beeinflussen können. Hierzu gehören

- das konkret betrachtete mathematische Inhaltsfeld (Funktionenlehre, Geometrie, etc.),
- die betrachtete prozessbezogene mathematische Tätigkeit sowie
- die Beschaffenheit des jeweiligen Items, etwa Antwortformat, Position im Fragebogen, allgemeine Schwierigkeit, etc.

Es gibt kaum Studien, die geschlechtsspezifische Effekte bis auf Itemebene transparent machen, nicht zuletzt sicher auch, da dies häufig als Unzulänglichkeit des entsprechenden Instruments begriffen wird (s.o.). I. d. R. werden geschlechtsspezifische Differenzen von Mittelwerten höchstens bis auf einzelne Themenfelder oder mathematische Tätigkeiten ausgewiesen. Diese lassen sich grob den zwei Haupt-Kompetenzdimensionen der deutschen Bildungsstandards zuordnen: *mathematischen Leitideen* bzw. *allgemeine* (oder auch *prozessbezogene*) *mathematische Kompetenzen* (KMK, 2015). Hierbei ist allerdings zu berücksichtigen, dass diese Zuordnung lediglich in ungefährer Form erfolgen kann, schließlich wurden entsprechende Studien in unterschiedlichen Bildungssystemen durchgeführt und orientieren sich – falls überhaupt – an entsprechend vielfältigen Lehrplänen.

Betrachtet man die Sachlage zu einzelnen mathematischen Themenfeldern, orientiert sich also zunächst an genannten Leitideen, fällt auf, dass Hyde et al. (1990, S. 147) in ihrer Meta-Analyse die größten Effektstärken zu Gunsten des männlichen Geschlechts im Bereich Geometrie feststellen. Diese beträgt durchschnittlich $d = 0.13$ und wurde anhand von 19 Einzeleffektstärken zusammengefasst. Auch im Rahmen einer Reanalyse der Daten aus PISA 2003 von Liu et al. (2008) zeigt sich der Inhaltsbereich „Raum und Form“ als am deutlichsten von geschlechtsspezifischen Abweichungen zu Gunsten der Jungen betroffen. Im Widerspruch hierzu steht, dass ebenfalls im Teilbereich Geometrie der 1995 durchgeführten TIMSS-Oberstufenerhebung geschlechtsspezifische Effekte am geringsten ausfallen. Diese betragen im Leistungskurs $d = 0.06$ und im Grundkurs $d = 0.02$ und sind somit zu vernachlässigen (vgl. Köller & Klieme, 2000, S. 401 f.). Ein weiterer Widerspruch ergibt sich auch innerhalb Deutschlands: So stellt sich im Rahmen des IQB-Ländervergleichs der Inhaltsbereich „Raum und Form“ als jene inhaltsbezogene Kompetenz heraus, welche am geringsten durch geschlechtsspezifische Differenzen belastet ist. Dieses Resultat zeigt sich sowohl in den Erhebungen 2012 für den neunten als auch bereits 2011 für den vierten Jahrgang gleichermaßen (vgl. Schroeders et al., 2013, S. 259; Böhme & Roppelt, 2012, S. 181). Eine entsprechende Übersicht über alle Leitideen ist für den IQB-Ländervergleich 2012 in Tabelle 2 dargestellt. Die abgebildeten Differenzen beziehen sich dabei jeweils auf die Differenzen der Mittelwerte für beide Geschlechter auf der Skala der Studie und werden für Gymnasien und sonstige Schulformen getrennt ausgewiesen.

Der für diese Abhandlung besonders interessante Bereich „funktionaler Zusammenhang“ befindet sich mit einer Effektstärke von $d = 0.25$ bzw. $d = 0.20$ im Mittelfeld der festgestellten geschlechtsspezifischen

Effekte. Im Übrigen kann auch hier wieder festgehalten werden, dass das Gymnasium für alle Leitideen die stärksten Differenzen zwischen den Geschlechtern zeigt.

Auch innerhalb der TIMSS-Erhebung von 1995 stellen sich die sog. Sachgebiete „Zahlen, Gleichungen und Funktionen“ sowie „Analysis“ mit $d = 0.29$ bzw. $d = 0.26$ für Leistungskurse als relativ geschlechtssensitiv heraus. Betrachtet man hingegen Grundkurse, reduzieren sich diese Werte auf $d = 0.12$ respektive $d = 0.04$. Ein ähnlicher Zusammenhang war bereits in Tabelle 1 für die Gesamtskala zu beobachten. Für den Bereich „Calculus“ ermitteln auch Hyde et al. (1990, S. 147) eine durchschnittliche Effektstärke von $d = 0.20$. Da dieser Wert jedoch auf der geringen Anzahl von lediglich zwei Einzeleffektstärken beruht, stellt sich für ihn keine Signifikanz ein.

Löst man sich von konkreten mathematischen Inhaltsbereichen und nimmt eher querliegende allgemeine mathematische Kompetenzen in den Blick, ergibt sich ein weniger widersprüchliches Bild: Hyde et al. (1990) stellen fest, dass gerade rechenintensive und kalkülhaltige Tests häufig zu Gunsten des weiblichen Geschlechts ausfallen. So bestimmen sie ausgehend von 45 Einzelgrößen eine mittlere Effektstärke von $d = -0.14$, die auf einen signifikanten geschlechtsspezifischen Effekt für den kognitiven Anforderungsbereich „computation“ deutet. Das männliche Geschlecht ist nach ihren Ergebnissen hingegen im Bereich „problem solving“ mit $d = 0.08$ – berechnet anhand von 48 Einzeleffekten – leicht aber signifikant im Vorteil. Betrachtet man lediglich die Gruppe der Fünfzehn- bis Achtzehnjährigen bzw. Neunzehn- bis Fünfundzwanzigjährigen, spitzt sich die Effektstärke auf Werte von $d = 0.29$ respektive $d = 0.32$ zu (vgl. Hyde et al., 1990, S. 147 f.).

Dieses Ergebnis steht im Einklang mit den Daten der TIMS-Studie von 1995. Hier zeigt sich eine Effektstärke von $d = -0.11$ in Grundkursen im Bereich „Routineverfahren“, während sich in Leistungskursen mit $d = 0.11$ erneut ein Vorteil für Männer äußert. In den Bereichen „Komplexe Verfahren“ und „Anwenden/Problemlösen“ ergeben sich zudem in Grundkursen leichte ($d = 0.14$ bzw. $d = 0.08$) und in Leistungskursen deutlichere ($d = 0.32$ bzw. $d = 0.30$) Effekte zu Gunsten des männlichen Geschlechts (vgl. Köller & Klieme 2000, S. 402). Dass Schülerinnen gerade in rechenlastigen oder Routineaufgaben nur leicht oder gar nicht benachteiligt sind und Schüler gerade im Bereich Problemlösen signifikante Leistungsvorsprünge aufweisen, zeigen auch andere Studien (z. B. Stewart et al., 2017; Spencer et al., 1999; Halpern & Wright, 1996; Harris & Carlton, 1993). Insgesamt kommen auch Köller und Klieme zu dem Schluss, „dass Aufgaben, die lediglich Routineverfahren zu ihrer Lösung erfordern, kleine oder keine Geschlechtsdifferenzen aufweisen“ (Köller & Klieme, 2000, S. 402).

Nicht zuletzt scheint die Stärke geschlechtsspezifischer Effekte auf Test- oder Itemebene auch durch mitunter äußere Aspekte der Itemgestaltung beeinflusst zu werden. So stellen Harris und Carlton (1993) fest, dass anwendungsbezogene Aufgaben, insbesondere Textaufgaben, signifikant leichter für männliche Probanden zu lösen sind. Während diese sich zudem nicht durch weiblich- bzw. männlich-stereotypisierte Einkleidungen beeinflussen lassen, schnitten Mädchen in einer Studie von Zohar und Gershikov (2008) in Aufgaben mit typisch männlichen Kontexten signifikant schlechter ab. Es lässt sich außerdem feststellen, dass Aufgaben, die Skizzen, Graphen oder Tabellen beinhalteten für Testteilnehmerinnen schwerer zu lösen sind (Ryan & Chiu 2001; Harris & Carlton 1993).

Weitere relevante gestalterische Elemente von Aufgaben betreffen etwa das vorgegebene Antwortformat: Innerhalb der Literatur zeichnet sich für Multiple-Choice- und Closed-Response-Aufgaben ein Vorteil für männliche Probanden ab (Le, 2009; DeMars, 1998), während DeMars (1998) für weibliche bei Constructed-Response-Formaten vernachlässigbare oder sogar umgekehrte Effekte feststellt. Dieser Zusammenhang gilt besonders für Probandinnen und Probanden des Spitzenfelds der verwendeten Fähigkeitsskala.
Dieser Effekt lässt sich jedoch auch unabhängig vom jeweiligen Antwortformat verallgemeinern: So lässt sich feststellen, dass geschlechtsspezifische

Leitidee	Gymnasium			Sonstige		
	Dif.	SE	d	Dif.	SE	d
Zahl	25	2.6	0.34	22	3.1	0.27
Messen	21	2.8	0.26	20	3.1	0.24
Raum und Form	10	2.6	0.13	8	3.3	0.10
Funktionaler Zusammenhang	18	2.5	0.25	16	3.0	0.20
Daten und Zufall	23	2.6	0.31	24	3.2	0.29

Tab. 2: Geschlechtsspezifische Differenzen nach Leitideen (entnommen aus Schroeders et al., 2013, S. 264)

Differenzen gerade für die jeweils best-performanten Probandinnen und Probanden besonders groß sind, während sich dieser Effekt umkehrt, wenn die anteilig jeweils schlechtesten Testteilnehmerinnen und -teilnehmer betrachtet werden (z. B. Contini et al., 2017; Fryer & Levitt, 2010; Ellison & Swanson, 2010; Penner & Paret, 2008; Penner, 2003; Xie & Shauman, 2003; Hedges & Novell, 1995).

3.4 Mögliche Ursachen der Differenzen

Die Ursachen der in den zahlreichen Studien ermittelten geschlechtsspezifischen Unterschiede sind nicht abschließend geklärt. Tatsächlich gibt es eine Vielzahl unterschiedlicher Erklärungsmodelle, welche z. B. in einer Publikation von Köller und Klieme (2000) zusammengefasst werden. Die entsprechende Kategorisierung solcher Erklärungsansätze scheint dabei auch nach Sichtung seither erschienener einschlägiger Veröffentlichungen aktuell: Köller und Klieme gruppieren Studien zur Ursache entsprechender geschlechtsspezifischer Differenzen in insgesamt vier Kategorien: biologische Ansätze, kognitive Ansätze, psychosoziale Modelle sowie Unterrichtsmodelle (vgl. Köller & Klieme, 2000, S. 376; s. auch Fox et al., 1977).

Hierbei sind unter der ersteren Kategorie der *biologischen Erklärungsansätze* z. B. evolutionspsychologische Ansätze zu verstehen, „die zum Beispiel annehmen, dass unterschiedlicher Selektionsdruck bei Männern und Frauen zu Differenzen in kognitiven Fähigkeiten, insbesondere in der Raumvorstellung, geführt habe“ (Köller & Klieme, 2000). Eine entsprechende Studie stammt etwa von Geary (1996). Weitere Studien machen etwa geschlechtsspezifische zerebrale Hemisphärenasymmetrien, hervorgerufen durch chromosomale Unterschiede (XX vs. XY) verantwortlich (z. B. Crow, 1994) oder liefern endokrine Erklärungsansätze wie etwa Hormonschwankungen im weiblichen Zyklus (z. B. Geary, 1989). Caplan und Caplan (2005) kritisieren den Aufwand hinsichtlich seiner ökonomischen und zeitlichen Dimensionen, der betrieben wird, um biologische Erklärungsansätze für die beobachtbaren Differenzen zu finden. Nicht zuletzt würden biologische Ansätze auch dadurch konterkariert, dass sich einzelne Länder finden ließen, in denen keine oder wenig geschlechtsspezifische Effekte in mathematischer Leistung bestehen (vgl. Caplan & Caplan, 2005, S. 25 f.). Es lassen sich zudem kaum neuere Studien finden, die einen ggfs. alleinigen biologischen Erklärungsansatz für die beobachtbaren Differenzen favorisieren.

Kognitive Ansätze hingegen nehmen nicht per se eine genetische Ursache der beobachteten Differenzen an und unterscheiden sich entsprechend von evolutionsbiologischen Ansätzen. Zu dieser Kategorie zählen beispielsweise Studien, die Raumvorstellung als Mediatorvariable zwischen den Variablen Geschlecht und Mathematikleistung annehmen. So kommen einige Studien zu dem Ergebnis, dass sich der Leistungsvorsprung des männlichen Geschlechts vor allem in Bereichen zeigt, die in besonderem Maße figurales oder räumliches Vorstellungsvermögen erfordern. Die entsprechende Hypothese einer mediierenden Raumvorstellungsvariable geht auf Sherman (1967) zurück und ging unter der Bezeichnung *Spatial Mediation Hypothesis* nach erstmaliger Verwendung durch Burnett et al. (1979) in die Literatur ein (vgl. Klieme, 1986, S. 136). So verlieren geschlechtsspezifische Leistungsvorsprünge in einigen Fällen ihre Signifikanz, wenn Raumvorstellungsfähigkeiten als Kovariate kontrolliert werden (vgl. Klieme 1986, S. 136). Es lassen sich weitere Studien finden, die Raumvorstellung – oder zumindest entsprechende Teilkompetenzen dieser – als geschlechtsspezifischen Mediator von Mathematikleistung identifizieren (z. B. Büchter, 2010; Grüßing, 2012; Voyer, 1996, 1998).

Psychosoziale Modelle hingegen stellen z. B. die häusliche Umwelt, durch die bereits eine frühe Festlegung auf Geschlechtsstereotype erfolgt, in den Mittelpunkt. Dies wirkt sich z. B. nach Eccles et al. (1990) negativ hinsichtlich der Einstellungen und Auseinandersetzung mit Mathematik von Mädchen aus. Insbesondere sog. „stereotype threat“ kann dazu führen, dass Mädchen bzw. Frauen ihre Kompetenzen bei der Bearbeitung einer mathematischen Leistungstestaufgabe nicht gänzlich ausschöpfen. So fielen die Resultate von Testteilnehmerinnen in einer Studie von Spencer et al. (1999) signifikant schlechter aus, wenn diese zuvor darauf aufmerksam gemacht wurden, dass Männer die vorgelegten Aufgaben im Mittel deutlich besser bearbeiten würden (s. auch Schwery et al., 2016; Good et al., 2008; Muzzatti & Agnoli, 2007; Quinn & Spencer, 2001; Walsh et al., 1999). Stoet und Geary (2012) weisen jedoch auf teilweise bestehende methodische Mängel entsprechender Studien hin und warnen vor der Überinterpretation der Bedeutung dieses Erklärungsansatzes.

Weitere Arbeiten gehen zudem von erhöhter mathematikbezogener Angst oder allgemeiner Testangst (z. B. Ganley & Vasilyeva, 2014; Cheema & Galluzzo, 2013; Chipman et al., 1992) oder einer generell niedrigeren Selbstwirksamkeitserwartung bei gleichem mathematischen Leistungsniveau bei Schülerinnen aus (z. B. Schwery et al., 2016; Cheema & Galluzzo, 2013; Else-Quest et al., 2010, 2013; Marsh & Yeung, 1998; Pajares, 1996; Wigfield & Eccles, 1992). Guiso et al. (2008) stellten zudem anhand einer Analyse von PISA-Daten fest, dass geschlechtsspezifische Leistungsdifferenzen in Mathematik

gerade in Ländern mit einer geschlechtergerechteren Kultur geringer ausfallen oder sogar verschwinden.

Die letzte Kategorie der *Unterrichtsmodelle* begründet die beobachteten geschlechtsspezifischen Leistungsdifferenzen beispielsweise durch eine unterschiedliche Behandlung beider Geschlechter durch die Lehrkraft, welche auf vorhandene Geschlechterstereotype auf Seiten der Lehrerin bzw. des Lehrers selbst zurückzuführen ist (z. B. Robinson-Cimpian et al., 2014; Gunderson et al., 2012; Fennema et al., 1990; Fennema & Peterson, 1987). Ferner wird davon ausgegangen, dass Curricula und Schulbücher sich meist an den Interessen und Lebenswelten von Jungen orientieren und somit Schülerinnen benachteiligen (z. B. Chipman et al., 1991).

Köller und Klieme (2000, S. 377) betonen aber auch, dass sich für die meisten Ansätze auch Arbeiten finden lassen, welche zu widersprüchlichen Resultaten kommen. Für Büchter (2010) scheint es daher plausibel, „dass Geschlechterunterschiede nur mit einem komplexen Wirkungsgefüge aus allen genannten Bereichen (empirisch wie theoretisch) erklärt werden können“ (Büchter, 2010, S. 56). Tatsächlich lassen sich auch Studien finden, die gerade ein Wechselgefüge einzelner Facetten der vorgestellten vier Kategorien untersuchen, etwa Zusammenhänge zwischen Raumvorstellung und „stereotype threat“ (z. B. Neuburger et al., 2012) oder Raumvorstellung und mathematikbezogener Angst (Ganley & Vasilyeva, 2014). Halpern et al. (2005) schlagen entsprechend ein „psychobiosoziales“ Modell als holistischeren Erklärungsansatz für die beobachtbaren mathematischen Leistungsdifferenzen zwischen den Geschlechtern vor.

3. Untersuchungsgegenstand und Forschungsfrage

3.1 Der FALKE-Test zum Funktionalen Denken

Das FALKE-Testinstrument (**F**unktionales **D**enken und Analysis: **L**ernen von **K**onzepten in der **E**inführungsphase) ist für einen Einsatz innerhalb des ersten Oberstufenjahres – der sog. *Einführungsphase* – konzipiert. Es fokussiert vor allem inhaltliches Verständnis im Bereich der Funktionenlehre und der frühen Analysis. Das Instrument besteht dabei aus zwei Teiltests: Während der erste Test Grundlagen der Funktionenlehre der Sekundarstufe I prüft und zu Beginn der Oberstufe eingesetzt werden sollte, fokussiert der zweite Test das bereits abgeschlossene erste Oberstufenjahr. In diesem wird typischerweise der Ableitungsbegriff erstmalig eingeführt, so dass dieser den Schwerpunkt des zweiten Tests bildet. Daneben wird der verständige Umgang mit Funktionen getestet, wie er in der Analysis notwendig ist. Hierzu gehört etwa der Umgang mit Parametern und Transformationen. Das Instrument stellt somit Funktionales Denken beim Übergang von Funktionenlehre zur Analysis in den Mittelpunkt.

Die Testitems orientieren sich dabei anhand des innerhalb der Dissertation von Klinger (2018) entwickelten Klassifikationsmodells. Dieses ist in Abbildung 2 dargestellt (vgl. auch Klinger & Barzel, 2018b). Es soll vor allem die Repräsentativität des verwendeten Itemsets sichern, so dass theoretisch relevante Aspekte des Funktionalen Denkens gleichmäßig durch einzelne Items vertreten werden.

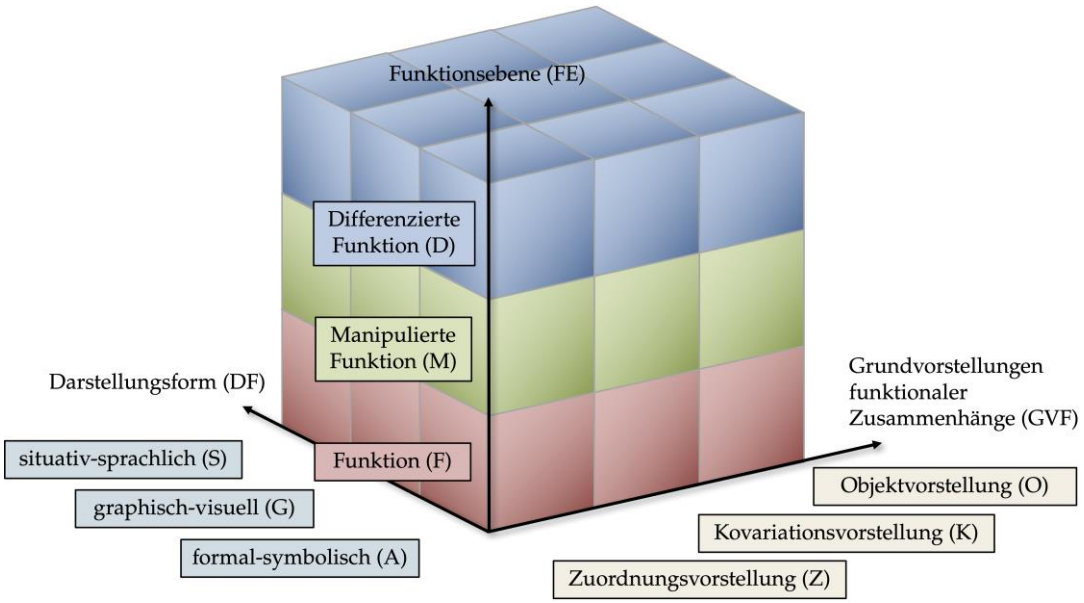

Abb. 2: Verwendetes Aufgabenraster des FALKE-Tests (Klinger, 2018)

Hierbei orientiert sich das Raster einerseits an der verwendeten Darstellungsform (DF). Hierzu gehört die situativ-sprachliche, die graphisch-visuelle sowie die formal-symbolische Darstellung eines funktionalen Zusammenhangs (z. B. Janvier, 1978; Swan, 1985).[3] Auf die Darstellungsform Tabelle wurde dabei vor dem Hintergrund einer besseren Testzeitökonomik verzichtet (s. auch Klinger 2018, Abschnitt 7.2). Gerade der Darstellungsvernetzung bzw. dem Wechsel zwischen unterschiedlichen Darstellungsformen kommt dabei eine besondere Rolle beim Lernen mathematischer Inhalte (z. B. Duval 2006), aber auch im Besonderen der Entwicklung von Verständnis innerhalb der Funktionenlehre bei (z. B. Swan, 1982; Laakmann, 2013).

Andererseits ist die in der jeweiligen Aufgabe im Vordergrund stehende Grundvorstellung funktionaler Zusammenhänge von Bedeutung (Vollrath, 1989; Malle, 2000). Hinsichtlich der Relevanz beider Kategorien für das verständige Durchdringen der Funktionenlehre und der Analysis besteht dabei breiter fachdidaktischer Konsens (z. B. Leuders & Prediger, 2005; Büchter, 2008; vom Hofe et al., 2015; Greefrath et al., 2016a).

Als dritte Dimension werden die entsprechenden Items anhand der auftretenen Funktionsebene unterschieden. Hiermit ist einerseits gemeint, dass Lernende flexibel zwischen der *üblichen Funktionsebene* (F) sowie ggfs. der Ebene der *differenzierten Funktion* (D) (Hahn, 2008; Hahn & Prediger, 2008) aber auch der Ebene einer *manipulierten Funktion* (M) wechseln müssen. Letztere umfasst dabei eine operative Veränderung einer Ausgangsfunktion, die nicht durch Differentiation, sondern durch das Wirken einer Transformation oder direkte Manipulation der entsprechenden Funktionsparameter der Term-Darstellung geschieht (vgl. Klinger, 2018, S. 119 ff.). Die Ebene der differenzierten Funktion wurde bei der Konzeption des FALKE-Testinstruments zusätzlich noch hinsichtlich der unterschiedlichen Grundvorstellungen des Ableitungsbegriffs konkretisiert, so dass das Itemset auch diesen Aspekt umfänglich abdeckt. Die entsprechend berücksichtigten Grundvorstellungen sind die *Änderungsratenvorstellung*, die *Tangentensteigungsvorstellung* sowie die *lokale Linearisierungsvorstellung* (Greefrath et al., 2016b; Danackwerts & Vogel, 2006; Blum & Törner, 1983).

Aufgrund seiner Anlage als Test zu Beginn der Einführungsphase umfasst der erste Test lediglich die Funktionsebenen F und M. Der zweite Test umfasst hingegen alle drei Ebenen (vgl. Klinger, 2018, S. 417 f.). Nahezu alle Items machen die Verwendung mindestens zweier Darstellungsformen notwednig, so dass i. d. R. Darstellungswechsel erforderlich sind, um Aufgaben erfolgreich zu bearbeiten. Darüber hinaus stehen stets unterschiedliche Grundvorstellungen (zu funktionalen Zusammenhängen bzw. zur Ableitungsfunktion) im Fokus.

Neben den durch obige Aufgabenklassifikation implizit gesetzten Kriterien für Test-Items der FALKE-Tests kommen weitere fachdidaktische Elemente zum Einsatz, um einen verständigen Umgang mit (Ableitungs-)Funktionen als Schlüsselfähigkeit der erfolgreichen Testbearbeitung zu gewährleisten. Hierzu zählt etwa der Einsatz sog. *qualitativer Funktionen*, bei denen man durch den Verzicht auf konkrete Zahlengrößen innerhalb der Aufgabenstellungen den Einsatz unreflektierter Kalkülfertigkeiten zur Bearbeitung versucht entgegenzuwirken (Klinger, 2018, S. 77 ff.; Stellmacher, 1986; Krabbendam, 1982). Dies geschieht etwa mit Aufgaben zu Füllprozessen, wie Abbildung 1 bereits exemplarisch zeigt. Darüber hinaus begünstigen einige Items häufige Fehler beim Umgang mit Funktionen, um somit auch diesen etwaig zugrunde liegende Fehlvorstellungen sichtbar zu machen (s. hierzu auch Nitsch 2015). Hierzu gehört einerseits der Graph-als-Bild-Fehler, bei dem Funktionsgraphen als „fotografische Abbilder von Realsituationen angesehen werden“ (Schlöglhofer, 2000, S. 16). Andererseits die sog. Illusion of Linearity, bei der Lernende lineare Zusammenhänge übergeneralisieren und insbesondere dort unterstellen, wo dies nicht angebracht ist (De Bock et al., 2002, 2007). Auch hier bildet das eingangs in Abbildung 1 dargestellte Item ein Beispiel, wenn Schülerinnen und Schüler etwa einen linearen Füllgraphen skizzieren.

Im zweiten Test sichern zudem geforderte Tätigkeiten wie das graphische Differenzieren, bei welchem lediglich ausgehend von einem Funktionsgraphen jener der zugehörigen Ableitungsfunktion skizziert werden muss, eine weitere Basis zur kalkülfreien und verständnisorientierten Erhebung einschlägiger Schülerleistungen (Klinger, 2018, S. 127 ff.; Hußmann & Prediger, 2010).

3.2 Forschungsfragen

Bisher wurde gezeigt, dass sich geschlechtsspezifische Effekte regelmäßig einstellen, versucht man mathematische Leistung oder Fähigkeiten mithilfe von psychometrischen Tests zu operationalisieren. Die Stärke dieser Effekte hängt von vielfältigen Faktoren ab. Zu diesen gehören der getestete Jahrgang, Schulformen, mathematische Inhaltsgebiete oder verlangte mathematische Tätigkeiten sowie weitere äußere

Merkmale entsprechender Items. Die meisten Leistungsstudien messen größere mathematische Kompetenzbereiche, etwa „mathematical literacy“ (PISA) oder „mathematical knowledge“ (TIMSS) (vgl. Leder & Forgash, 2018, S. 691). Die FALKE-Erhebung bietet hier die Möglichkeit einen deutlich stärker mit einem mathematischen Inhaltsbereich verbundenen Test hinsichtlich seiner geschlechtsspezifischen Eigenschaften zu untersuchen. Hieraus sollen vor allem für den Bereich des Funktionalen Denkens Rückschlüsse auf die Bedeutung des Geschlechts geschlossen werden. Konkret soll hierzu die folgende Forschungsfrage gestellt und im weiteren Verlauf des Artikels beantwortet:

Welche Stärke geschlechtsspezifischer Effekte lässt sich für den Bereich „Funktionales Denken“ anhand des FALKE-Tests ausmachen und welche Item-Merkmale und inhaltlichen Teilanforderungen beeinflussen hierbei die Stärke solcher geschlechtsspezifischen Effekte?

Die Beantwortung dieser Frage soll daher vor allem explorativen Charakter haben und Grundlage weiterer Forschung in diesem Bereich sein.

3.3 Stichprobe

Die FALKE-Tests wurden zu Beginn bzw. gegen Ende des Schuljahres 2014/15 in Nordrhein-Westfalen großflächig eingesetzt. Hierzu wurden Lehrkräfte einerseits über ein Rundschreiben an alle Schulen mit gymnasialer Oberstufe, andererseits über die Fortbildungsreihe „GTR kompakt“ (Klinger et al. 2018) um Teilnahme gebeten. Für den ersten Test kam so eine Stichprobe von 3202 Schülerinnen und Schüler (50.0 % w., 49.5 % m.), für den zweiten Test von 2665 Schülerinnen und Schüler (48.9 % w., 50.3 % m.) zustande.

Hierbei umfassen beide Testzeitpunkte im Wesentlichen dieselben Schülerinnen und Schüler. Der erste Test wurde zu Beginn der Einführungsphase, also dem ersten Oberstufenjahr, der zweite Test gegen Ende der Einführungsphase ausgeführt. In Nordrhein-Westfalen lässt sich das Abitur im Wesentlichen an den drei Schulformen Gymnasium, Gesamtschule sowie an Beruflichen Gymnasien, welche in Berufskollegs integriert sind, ablegen.[4] Die Schülerinnen und Schüler verteilen sich dabei innerhalb der Stichprobe wie in Abbildung 3 dargestellt.

Die Testleitung übernahm die jeweiligen Lehrkraft der teilnehmenden Lerngruppe. Für die Bearbeitung

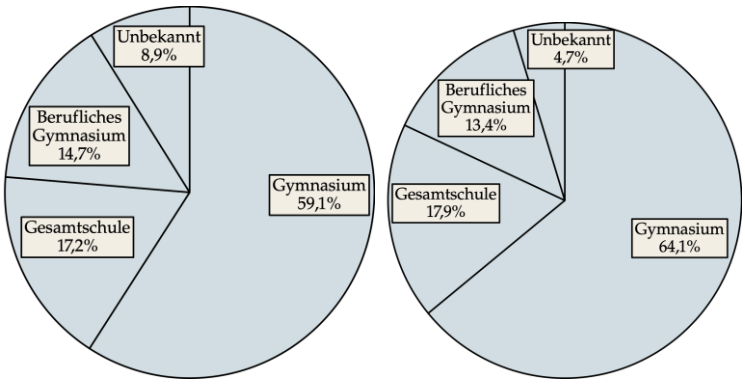

Abb. 3: Verteilung der Stichprobe hinsichtlich der drei Schulformen im ersten (links) bzw. zweiten Test (rechts)

	Erster Test					Zweiter Test				
	M_α	SD_α	M_p	SD_p	n	M_α	SD_α	M_p	SD_p	n
männlich	-0.33	1.19	8.49	2.75	1584	0.47	1.19	12.15	4.17	1340
weiblich	-0.77	1.18	7.44	2.75	1602	0.05	1.10	10.63	3.99	1304
gesamt	-0.55	1.21	7.96	2.81	3202	0.26	1.16	11.39	4.16	2665
Differenz	0.44	0.01	1.05	0.06	-18	0.42	0.09	1.52	0.18	36
d	0.37	—	0.38	—	—	0.37	—	0.37	—	—

M_α: Mittel der Fähigkeitsparameter, SD_α: Standardabweichung der Fähigkeitsparameter, M_p: Mittel der Anzahl gelöster Items, SD_p: Standardabweichung der Anzahl gelöster Items, n: Anzahl der Probanden innerhalb der jeweiligen Gruppe, d: Maß der Effektstärke nach Cohen (1988)

Tab. 3: Kennzahlen zu geschlechtsspezifischen Abweichungen im ersten und zweiten Test

standen den Schülerinnen und Schüler je Test 45 Minuten während des Unterrichts aber unter üblichen Prüfungsbedingungen zur Verfügung. Ferner wurden zwei itemgleiche Erhebungsbogenvarianten innerhalb jeder Lerngruppe eingesetzt. Die Zustellung entsprechender Bögen erfolgte jeweils postalisch. Die Korrektur und Datenerfassung erfolgte nicht durch die Lehrkräfte selbst, sondern von eigens geschulten studentischen Hilfsakräften sowie dem Studienautor. Parallel zum Test wurde ein einseitiger Fragebogen im Rahmen der Studie „GTR NRW“ administriert (s. Thurm et al., 2015), der jedoch für die vorliegende Studie keine weitere Bewandnis hat.

Aufgrund des oben beschriebenen Auswahlverfahrens kann nicht von einer Repräsentativität der Stichprobe für die gesamte Schülerschaft ausgegangen werden, auch wenn ein Vergleich mit offiziellen Daten des Kultusministeriums diese zumindest suggeriert. So stimmen Geschlechter- wie auch Schulformverhältnisse im Wesentlichen mit denen der zugrunde liegenden Population überein (vgl. Klinger, 2018, S. 211 f.).

Für die gezogene Stichprobe hat sich gezeigt, dass beide Tests eine Skalierung mit dem eindimensionalen Rasch-Modell (Rasch, 1980) zulassen (Klinger, 2018).

4. Geschlechtsspezifische Effekte im FALKE-Test

Bisher wurde bereits dargelegt, dass mathematische Leistungstests häufig eine geschlechtsspezifische Leistungsdisposition zu Gunsten der männlichen Probanden aufweisen. Hier hat sich gezeigt, dass diese Effekte bereits im Primarbereich sichtbar sind und über die Sekundarstufe I bis hin zur Oberstufe tendenziell zunehmen. Wie der IQB-Ländervergleich 2012 zeigt, ist Nordrhein-Westfalen insofern insbesondere betroffen, als dass sich hier die stärksten Effekte gegen Ende der Sekundarstufe I zeigen (vgl. Schroeders et al., 2013, S. 265). Hinzu kommt, dass der Bereich „Funktionaler Zusammenhang“ mit einer Effektstärke von $d = 0.25$ an Gymnasien auch inhaltlich betroffen ist (vgl. Schroeders et al., 2013, S. 259). Dies spiegelt sich auch in den etwas älteren Ergebnissen der TIMS-Studie von 1995 wider. Hier zeigen insbesondere Leistungskurse im 13. Schuljahr mit $d = 0.29$ bzw. $d = 0.26$ auffällige Effektstärken für die Bereiche „Zahlen, Gleichungen und Funktionen“ respektive „Analysis“ (vgl. Köller & Klieme 2000, S. 402).

Im Folgenden sollen daher die im Rahmen der FALKE-Erhebung gewonnenen Leistungstestdaten hinsichtlich beobachtbarer geschlechtsspezifischer Leistungsdifferenzen untersucht werden; dies zunächst auf globaler Ebene, d.h. über den gesamten Test hinweg. In Abschnitt 4.2 werden sodann, um mitunter auch etwaigen Simpson-Effekten vorzubeugen, die Daten hinsichtlich der konkreten Schulform analysiert. In Abschnitt 4.3 werden schließlich geschlechtsspezifische Effekte bis auf Itemebene verfolgt. Hierbei werden zudem Hypothesen bezüglich relevanter Aufgaben-Merkmale erarbeitet, die entsprechende Effekte besonders begünstigen oder unterdrücken.

4.1 Globale Testebene

Um zunächst die globale Testebene hinsichtlich der genannten Effekte zu untersuchen, sind in erster Instanz entsprechende Kennzahlen für die verwendeten Instrumente in Tabelle 3 dargestellt, welche auch die Effektstärke d nach Cohen (1988) beinhaltet.

Insgesamt haben an beiden Tests etwa gleichviele Mädchen und Jungen teilgenommen. Vergleicht man die Leistungen beider Geschlechter, zeigen sich auch für die vorliegende Arbeit und für beide Tests signifikante geschlechtsspezifische Abweichungen zu Gunsten der männlichen Testteilnehmer. Diese befinden sich mit Werten von $d = 0.37$ bzw. $d = 0.38$ jeweils im unteren mittleren Bereich (s. Bewertung nach Hyde, 2005 in Abschnitt 2.1).

Diese Werte erscheinen vor dem Hintergrund der innerhalb der TIMSS- und PISA-Erhebungen gefundenen Werte als erhöht (s. Tabelle 1). Gerade aber im Vergleich zu reinnationalen Studien wie dem IQB-Ländervergleich oder der KESS-Studie stellen sich durchaus ähnliche Werte ein. Ursachen hierfür dürften sich u.a. in einer unterschiedlichen Studienmethodik finden (s. Abschnitt 2.2).

Es lässt sich weiterhin feststellen, dass sich in beiden Tests dieser Studie im Wesentlichen gleiche Effektstärken einstellen. Dies ist insofern plausibel, als dass in beiden Tests eine wesentliche Anzahl der getesteten Personen identisch ist und zudem fünf Items in Form von Anker-Items übereinstimmen.

Im Mittel erhalten männliche Probanden je nach betrachtetem Test einen um 0.44 bzw. 0.42 erhöhten Fähigkeitsparameter im Rahmen der Rasch-Modellierung. Betrachtet man dieses Resultat auf Ebene der Items, zeigt sich für den ersten Test, dass Männer im Mittel 1.05 Items mehr lösen als Frauen. Für den zweiten Test erhöht sich dieser Wert auf 1.52 Items, wenngleich dieser angesichts der insgesamt etwas höheren Itemanzahl im zweiten Test zu relativieren ist. Die Standardabweichung schwankt für beide Teilerhebungen insgesamt in den geschlechtsspezifischen Gruppen nur unwesentlich.

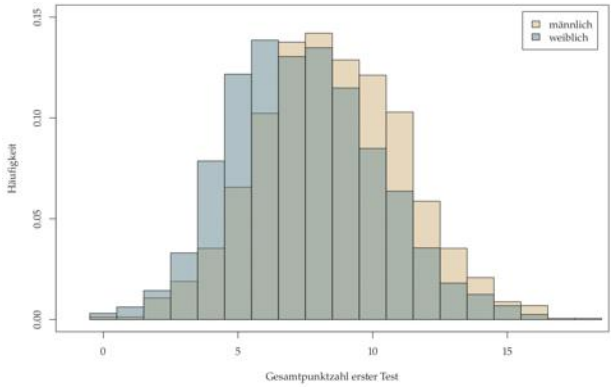

Abb. 4: Histogramm der Gesamtpunktzahl im ersten Test gruppiert nach Geschlecht

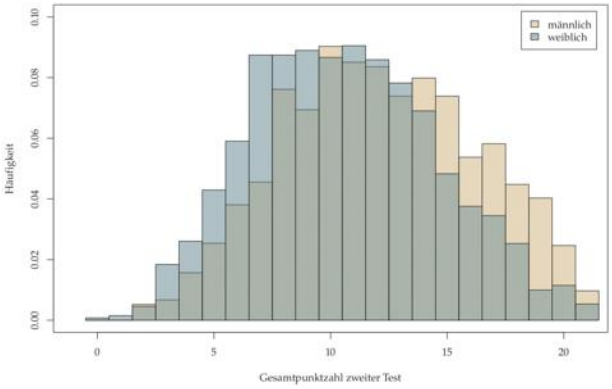

Abb. 5: Histogramm der Gesamtpunktzahl im zweiten Test gruppiert nach Geschlecht

Dass die in Tabelle 3 dargestellten Mittelwertunterschiede nicht etwa ausschließlich die Folge einer überproportionierten männlichen Spitzengruppe sind, wie innerhalb der Literatur häufig berichtet wird, sondern sich über große Teile der Verteilung erstrecken, zeigen Abbildung 4 und 5. Hierbei zeigt die erste Grafik die ineinander verschränkten Histogramme für beide Geschlechtergruppen im ersten Test, während die zweite Abbildung auf ähnliche Weise die Daten des zweiten Tests darstellt. Dabei wird die relative Häufigkeit den jeweiligen im Test erreichten Gesamtpunktzahlen gegenübergestellt. Diese ergeben sich als Anzahl der gelösten Items einer Testeilnehmerin bzw. eines Testteilnehmers. Zu beobachten ist dabei, dass gerade in der Mitte der Verteilung weibliche und männliche Probanden in etwa gleichauf sind.

Für den ersten Test ist deutlich zu erkennen, dass der Bereich niedriger Gesamtpunktzahlen zwischen null und sechs Punkten durch das weibliche Geschlecht dominiert wird. Am deutlichsten fällt dieser Effekt für die Punktzahlen 4 bis 6 aus. Hingegen sind ab einer Punktzahl von sieben Punkten Jungen im Vorteil, wobei dieser Effekt noch im mittleren Bereich und am Ende der Skala am geringsten ausfällt. Die Bereiche, in denen sich beide Verteilungen maximal

	Erster Test					Zweiter Test				
Gymnasium	**M_α**	**SD_α**	**M_p**	**SD_p**	**n**	**M_α**	**SD_α**	**M_p**	**SD_p**	**n**
männlich	-0.18	1.20	8.83	2.75	917	0.64	1.19	12.74	4.10	843
weiblich	-0.62	1.02	7.79	2.60	966	0.19	1.09	11.14	3.97	851
gesamt	-0.41	1.17	8.30	2.72	1883	0.42	1.16	11.94	4.11	1694
Differenz	0.44	0.18	1.04	0.15	-49	0.46	0.10	1.60	0.13	-8
d	0.38	—	0.39	—	—	0.40	—	0.40	—	—
Gesamtschule	**M_α**	**SD_α**	**M_p**	**SD_p**	**n**	**M_α**	**SD_α**	**M_p**	**SD_p**	**n**
männlich	-0.60	1.02	7.83	2.47	249	0.35	1.02	11.79	3.80	209
weiblich	-1.15	1.10	6.52	2.51	300	-0.27	0.92	9.47	3.49	263
gesamt	-0.90	1.10	8.30	2.58	549	0.01	1.01	10.50	3.81	472
Differenz	0.54	-0.08	1.31	-0.04	-51	0.62	0.10	3.32	0.31	-54
d	0.53	—	0.53	—	—	0.64	—	0.64	—	—
Berufskolleg	**M_α**	**SD_α**	**M_p**	**SD_p**	**n**	**M_α**	**SD_α**	**M_p**	**SD_p**	**n**
männlich	-0.69	1.18	7.67	2.77	274	-0.09	1.19	10.09	4.29	223
weiblich	-1.31	1.13	6.16	2.54	195	-0.50	1.03	8.62	3.75	130
gesamt	-0.95	1.20	7.04	2.77	469	-0.24	1.15	9.55	4.16	353
Differenz	0.63	0.04	1.51	0.23	79	0.41	0.17	1.47	0.54	93
d	0.57	—	0.57	—	—	0.37	—	0.36	—	—

M_α: Mittel der Fähigkeitsparameter, SD_α: Standardabweichung der Fähigkeitsparameter, M_p: Mittel der Anzahl gelöster Items, SD_p: Standardabweichung der Anzahl gelöster Items, n: Anzahl der Probanden innerhalb der jeweiligen Gruppe, d: Maß der Effektstärke nach Cohen (1988)

Tab. 4: Kennzahlen zu geschlechtsspezifischen Abweichungen im ersten und zweiten Test bezogen auf die Schulform

unterscheiden, haben dabei jeweils eine Breite von etwa einer Standardabweichung.

Für den zweiten Test ergibt sich im Vergleich hierzu ein leicht verzerrtes Bild. Insgesamt ist die Verteilung weniger symmetrisch. Bis zu einer Gesamtpunktzahl von 13 gelösten Items haben Mädchen einen höheren Anteil an den entsprechenden Punktzahlen. Eine Ausnahme bildet die Punktzahl 10, bei der eine leichte Überproportion des männlichen Geschlechts besteht. Die höheren Punktzahlen ab 14 Punkten werden erneut häufiger durch Probanden männlichen Geschlechts besetzt, so dass Personen mit sehr hoher Gesamtleistung überwiegend Männer sind.

4.2 Vergleich hinsichtlich Schulform

Im Rahmen der FALKE-Erhebung wurden Gymnasien, Gesamtschulen sowie Berufskollegs, die zur allgemeinen Hochschulreife führen, und somit insgesamt drei unterschiedliche nordrhein-westfälische Schulformen einbezogen. Da sich innerhalb der Literatur zeigt, dass geschlechtsspezifische Effekte durchaus über unterschiedliche schulische Milieus hinweg variieren können und zudem auch etwaige Simpson-Paradoxa mitgedacht werden sollten, wird auch für die vorliegende Studie eine entsprechende Analyse hinsichtlich der einbezogenen Schulformen durchgeführt. Tabelle 4 zeigt die hierzu notwendigen Kennzahlen und schlüsselt diese den unterschiedlichen Schulformen entsprechend auf.

Es zeigen sich innerhalb aller Schulformen und für beides Tests signifikante Effekte zu Gunsten männlicher Testteilnehmer. Diese fallen tendenziell stärker an Gesamtschulen und Berufskollegs aus als am Gymnasium. Dort zeigen sich die geringsten geschlechtsspezifischen Abweichungen. Eine Ausnahme bilden hierbei die Schülerinnen und Schüler des Berufskollegs im zweiten Test. Mit einer Effektstärke von d = 0.36 (in Bezug auf die Gesamtpunktzahl) fallen die Differenzen knapp weniger intensiv aus als am Gymnasium.

Während das Gymnasium in beiden Erhebungen einen in etwa gleichbleibenden Effekt zwischen d = 0.38 und d = 0.40 zeigt, schwankt die Effektstärke an Gesamtschulen und Berufskollegs zwischen beiden Tests. An Gesamtschulen fällt sie im ersten Test etwas schwächer aus als im zweiten. An Berufskollegs hingegen zeigt sich ein umgekehrter Effekt, so dass sich hier die Effektstärke reduziert. Dies kann jedoch mit der überproportional reduzierten Stichprobengröße zwischen erstem und zweitem Test für diese Schulform zusammenhängen. So bricht die Anzahl der erfassten Schülerinnen und Schüler von 469 für den ersten Test auf 353 für den zweiten Test deutlicher ein als für die anderen Schulformen.

Vergleicht man die Effektstärke geschlechtsspezifischer Differenzen für die Gesamtstichprobe aus Tabelle 3 mit den jeweils nach Schulform getrennten Gruppen, stellt sich auch für die FALKE-Testdaten ein gewisses Simpson-Paradoxon ein. So liegen alle Effektstärken innerhalb der schulformspezifischen Subgruppen wenigsten gleich auf, z. T. aber auch deutlich über die Effektstärke für die Gesamtstichprobe.

4.3 Vergleich auf Itemebene

Um schließlich einen genaueren Eindruck zu ermöglichen, worauf die beobachteten geschlechtsspezifischen Effekte aus fachdidaktischer Perspektive zurückgeführt werden können, sind die Lösungsquoten aller Items nach Größe der entsprechenden Differenzen in zwei Balkendiagrammen dargestellt.[5] Hierbei

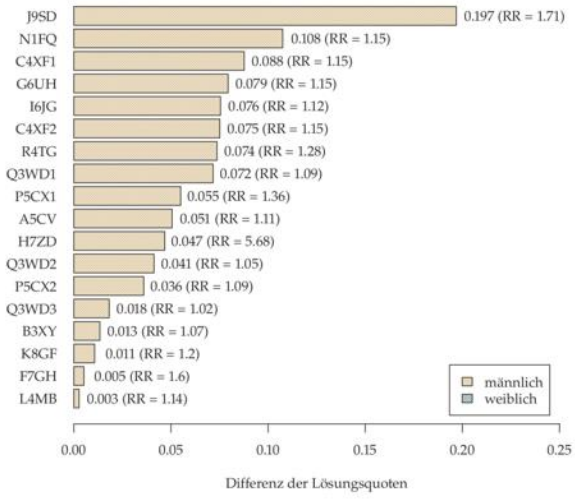

Abb. 6: Differenz der durchschnittlichen Lösungsquoten aller Items des ersten Tests für beide Geschlechter

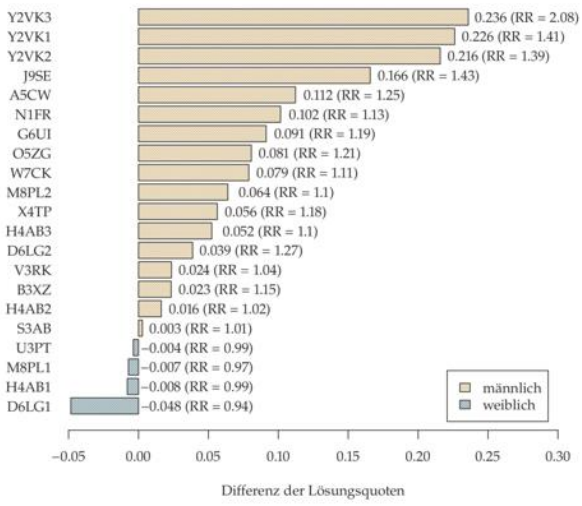

Abb. 7: Differenz der durchschnittlichen Lösungsquoten aller Items des zweiten Tests für beide Geschlechter

sind die Items des ersten Tests in Abbildung 6 abgetragen, jene für den zweiten Test in Abbildung 7. Positive Differenzen deuten auf einen Leistungsvorsprung der Männer, negative auf einen Leistungsvorsprung der Frauen, wobei nun wieder die Gesamtstichprobe an allen Schulformen zugrunde gelegt wird.

Weiterhin ist auch das sog. *relative Risiko RR* (engl. „relative risk" oder „risk ratio") neben den Balken dargestellt. Es soll neben der reinen Differenz der Lösungsquoten als weiteres Maß der Effektstärke dienen, da Cohens *d* sich im Ein-Item-Fall seiner Konstruktion nach nicht eignet (vgl. Borenstein et al., 2009, S. 33 ff.). Es wird als Quotient der mittleren Lösungsquote von Männern und jener von Frauen für jedes Item gebildet, so dass Werte größer als 1 einen Vorteil des männlichen, Werte kleiner als 1 einen Vorteil des weiblichen Geschlechts für das jeweilige Item wiedergeben.

Erster Test

Für den ersten Test sind alle Differenzen positiv, so dass sich für jedes der 18 Testitems eine höhere Lösungsquote für das männliche Geschlecht ergibt.

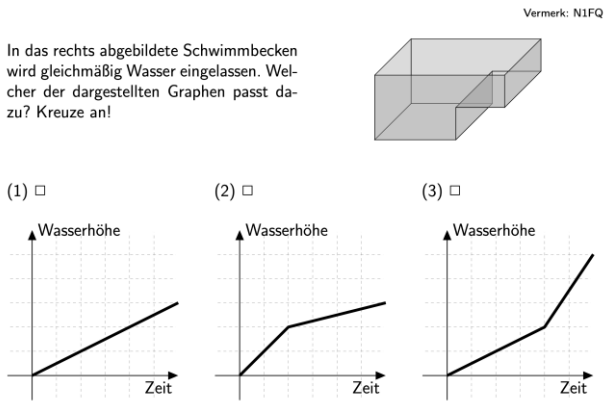

Abb. 8: Aufgabe „Schwimmbecken" (Kennung N1FQ) des FALKE-Tests (Klinger 2018)

Dabei zeigt das bereits dargestellte Item „Kegelfüllung" (J9SD) (Abbildung 1) die größte Differenz zwischen durchschnittlicher Lösungsquote der männlichen und durchschnittlicher Lösungsquote der weiblichen Probanden. Insgesamt wird die entsprechende Aufgabe von 27.8 % der weiblichen und von 47.5 % der männlichen Probanden gelöst, so dass sich insgesamt die dargestellte Differenz von 19.7 Prozentpunkten ergibt. Das Item weist zudem ein relatives Risiko von RR = 1.71 auf, d.h. Männer lösen es 1.71-mal häufiger als Frauen.

Die zweitgrößte Differenz zeigt sich bei dem insgesamt deutlich leichteren Item N1FQ und entspricht 10.8 Prozentpunkten. Das relative Risiko zeigt sich hingegen vergleichsweise nahe bei 1, was auf die insgesamt hohen Lösungsquoten für dieses Item zurückzuführen ist: Es wird von 84,5 % der männlichen und 73,7 % der weiblichen Testteilnehmer gelöst. Das Item ist in Abbildung 8 dargestellt.

In beiden Aufgaben lassen sich leicht inhaltliche Übereinstimmungen finden. So handelt es sich jeweils um Füllstandsaufgaben, bei denen eine gegebene Situation in einen Füllstand-Zeit-Graphen überführt werden muss. Aufgabe „Kegelfüllung" stellt dabei für beide Geschlechter die größere Herausforderung dar, sicherlich auch da hier im Gegensatz zu Item „Schwimmbecken" keine Antwortmöglichkeiten zur Verfügung stehen, sondern der Graph frei zu skizzieren ist.

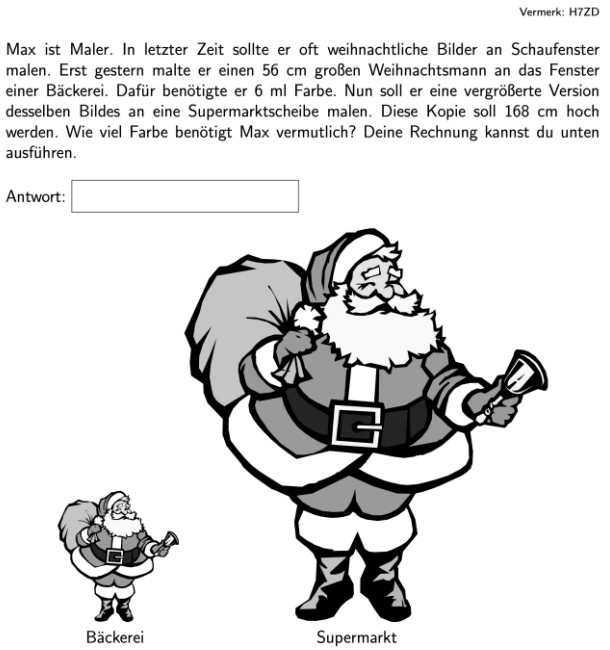

Abb. 9: Aufgabe „Weihnachtsmann" (Kennung H7ZD) des FALKE-Tests (Klinger 2018)

Zwar weniger hinsichtlich der beobachteten Differenz, jedoch hinsichtlich eines beobachteten relativen Risikos von RR = 5.68, ist auch Item H7ZD auffällig. Insgesamt weist die Aufgabe mit einer durchschnittlichen Lösungsquote von 3.3 % eine hohe empirische Schwierigkeit auf. Die korrekte Lösung lautet „54 ml". Sie wird von männlichen Probanden in 5.68 % der Fälle gelöst, von Mädchen hingegen lediglich in 1.00 % der Fälle. Die entsprechende Aufgabe ist Abbildung 9 zu entnehmen. Wie bereits Aufgabe „Kegelfüllung" zielt auch dieses Item auf eine mögliche Illusion of Linearity ab.

Jene vier Items mit der kleinsten festgestellten Differenz (B3XY, K8GF, F7GH, L4MB) bilden auch gleichzeitig die Gruppe der schwierigsten Testitems. Die geschlechtsspezifischen Effekte werden hier nicht signifikant. Item L4MB wird beispielsweise von 2.3 % der männlichen sowie von 2.0 % der weiblichen Probanden korrekt bearbeitet.

Zweiter Test

Im zweiten Test zeigt sich die in Abbildung 10 dargestellte Itemfamilie Y2VK mit allen Teilaufgaben als stark vom jeweiligen Geschlecht des bearbeitenden Probanden abhängig.

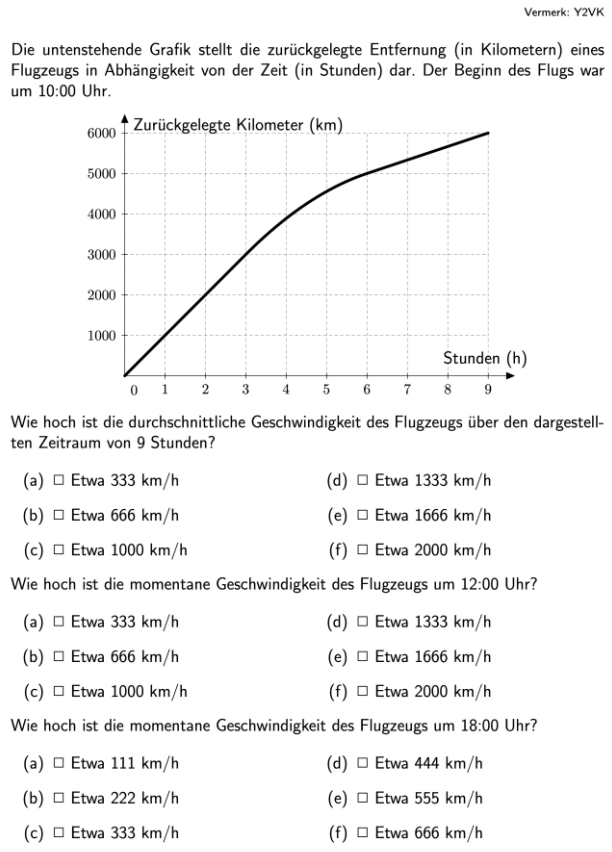

Abb. 10: Aufgabe „Flugzeug“ (Kennung Y2VK) des FALKE-Tests (Klinger 2018)

Diese besteht aus den drei als Einzelitems kodierten Teilaufgaben Y2VK1, Y2VK2, Y2VK3. Die Differenz beträgt für alle Items jeweils über 20 Prozentpunkte, was sich auch in teilweise hohen bis sehr hohen relativen Risiken zu Ungunsten des weiblichen Geschlechts niederschlägt. Im Rahmen dieser Items muss jeweils ausgehend von einem Weg-Zeit-Graphen die Durchschnitts- bzw. Momentangeschwindigkeit für ein Intervall bzw. einen spezifischen Zeitpunkt bestimmt werden. Es handelt sich somit im weiteren Sinne um Aufgaben im Themenfeld Differentialrechnung. Konkret ist insbesondere in der zweiten wie dritten Teilaufgabe die Vorstellung der Ableitungsfunktion (bzw. der Momentangeschwindigkeit) als lokale Steigung des Graphen bzw. der Tangente zur Bearbeitung hilfreich. In diesem Sinne kann das dargestellte Item als Aufgabe zur Tangenten- und Änderungsratenvorstellung des Ableitungsbegriffs aufgefasst werden (vgl. Klinger, 2018, S. 313 ff.).

Ebenfalls durch eine relativ große Differenz in den geschlechtsspezifischen Lösungsquoten gezeichnet, sind die Füllstand-Aufgaben J9SE und N1FR, welche bereits im ersten Test auffällig waren und im zweiten Test als Ankeritems mit leicht veränderten Zahlen erneut eingesetzt wurden.

Im Gegensatz zum ersten Test zeigen für den zweiten Test auch insgesamt vier Items Differenzen zu Gunsten des weiblichen Geschlechts, welche jedoch mit Ausnahme von Item D6LG1 kein Signifikanzniveau erreichen. Letzteres ist in Abbildung 11 dargestellt. Es ist mit einer durchschnittlichen Lösungsquote von 83.3 % sicherlich auch aufgrund einer Ratewahrscheinlichkeit von 50 % eines der einfachsten Items im Test. Hierbei wird es von 85.8 % der weiblichen und von 81.0 % der männlichen Testteilnehmer korrekt beantwortet. Neben einer Argumentation über das Vorzeichen des Führungskoeffizienten der Ableitungsfunktion, lässt sich die Antwort auch über graphisches Differenzieren begründen.

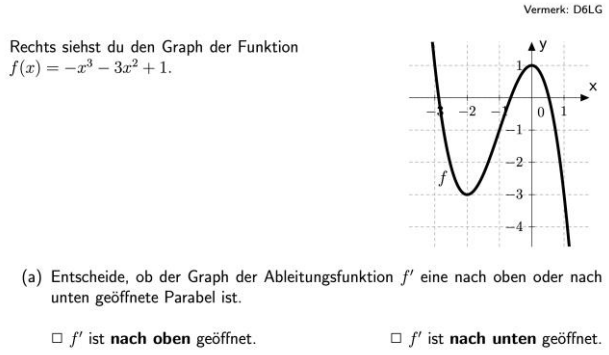

Abb. 11: Aufgabe „Parabelöffnung“ (Kennung D6LG1) des FALKE-Tests (Klinger 2018)

Insgesamt ist auffällig, dass Items, die durch graphisches Differenzieren zu lösen sind, sich in Abbildung 7 jeweils kaum oder in sogar umgekehrte Richtung geschlechtsspezifische Effekte zeigen (Kennungen U3PT, S3AB und V3RK). Neben diesen zeigen sich zudem auch Items, bei denen symbolisch, d. h. mithilfe der Ableitungsregeln und ausschließlich auf Term-Ebene, differenziert werden muss im entsprechend unauffälligen Bereich (H4AB1, H4AB2 und H4AB3).

Gesamtschau

Sucht man insgesamt nach Gemeinsamkeit in den vorliegenden Daten, so zeichnet sich ab, dass Probandinnen einerseits bei Aufgaben zu Füllprozessen, die in Form eines Funktionsgraphen visualisiert werden müssen, benachteiligt sind. So zeigen die entsprechenden Aufgaben „Kegelfüllung“ und „Schwimmbecken“ in beiden Tests entsprechende Auffälligkeiten. Dies kann vor dem Hintergrund der in Abschnitt 3.4 dargestellten Literatur möglicherweise über eine besondere Beanspruchung des räumlichen Vorstellungsvermögens durch entsprechende Aufgaben erklärt werden. In beiden Items müssen Probanden die

äußere Form eines Behältnisses im Kontext des Befülltwerdens auf die Form eines Funktionsgraphen übertragen.

Eine weitere Gemeinsamkeit der Items „Kegelfüllung“ und „Schwimmbecken“ findet sich ebenfalls im dem Item zugrunde liegenden funktionalen Prozess. So muss in beiden Fällen auf einen nicht-linearen Zusammenhang geschlossen werden. Hierbei spielt die bereits beschriebene Illusion of Linearity (De Bock et al., 2007) eine besondere Rolle. Diese ist zudem auch relevant im Item „Weihnachtsmann“, welches in Form eines hohen relativen Risikos mit geschlechtsspezifischen Effekten belastet ist. Ein genauerer Blick auf die durch Schülerinnen und Schüler gegebenen Antworten, zeigt jedoch, dass Mädchen die auf eine Übergeneralisierung linearer Zusammenhänge deutende Antwort „18 ml“ nicht signifikant häufiger geben als Jungen (1214 von 1584 für m; 1203 von 1602 für w; $p > .05$ nach exaktem Test nach Fisher).

Im zweiten Test zeigen sich vor allem Aufgaben zur Durchschnitts- und Momentangeschwindigkeit auffällig. Das einheitliche Bild innerhalb von drei Aufgaben lässt vermuten, dass Mädchen auch in diesem Bereich gegenüber Jungen benachteiligt sind.

Als weniger auffällig zeigen sich im ersten Test solche Items, die von besonderer Schwierigkeit sind. Möglicherweise würden diese bei einer noch größeren Stichprobe an Signifikanz gewinnen. Im zweiten Test zeigen sich gerade solche Aufgaben als wenig auffällig (oder sogar auffällig zu Gunsten des weiblichen Geschlechts), welche eher prozedurales Wissen erfordern und sich durch einstudierte Kalküle lösen lassen. Zu diesen gehören Items zum symbolischen Differenzieren (H4AB1, H4AB2 und H4AB3) sowie bedingt das Item „Parabelöffnung“ (D6LG1), welches sich ebenfalls mittels Ableitungskalkül aber auch durch graphisches Ableiten erfolgreich bearbeiten lässt.

Auffällig ist auch, dass alle Items zum graphischen Ableiten schwache geschlechtsspezifische Effekte aufweisen. Neben Item D6LG1 fokussieren insbesondere die Aufgaben U3PT, S3AB sowie V3RK das graphische Differenzieren. Für alle Items lassen sich keine signifikanten Geschlechtereffekte zu Gunsten der Jungen feststellen.

Das sich abzeichnende Bild lässt sich dabei wie folgt interpretieren: Die Ausgeglichenheit der Leistungen von Schülerinnen und Schülern im Bereich des graphischen Ableiten kann vor dem Hintergrund der literaturbekannten relativen Stärke der Frauen bei kalkülhaltigen Aufgaben (Hyde et al., 1990) als Indiz dafür gewertet werden, dass auch Aufgaben zum graphischen Differenzieren vor allem über einen einstudierten Kalkül bearbeitet werden. Hierbei stehen dann weniger die Vorstellungen zum Ableitungsbegriff im Fokus als eher das einstudierte Zuordnen charakteristischer Punkte sowie die Identifikation des Vorzeichens der Funktionswerte zwischen diesen (vgl. Klinger, 2018, S. 127 ff.). Dass gerade Item D6LG1 als einzige Aufgabe sichtbar Frauen bevorteilt, kann dann so erklärt werden, dass es eine größere Auswahl an unterschiedlichen Verfahren zulässt, die zu einer erfolgreichen Bearbeitung führen. Deutlich zu betonen ist hierbei jedoch der spekulative Charakter vorgenannter Überlegungen.

5. Fazit

Im theoretischen Teil dieses Beitrags wurden geschlechtsspezifische Abweichungen in mathematischen Leistungstests bereits dargelegt. Diese lassen sich regelmäßig beobachten und fallen in aller Regel zu Gunsten des männlichen Geschlechts aus. Die Stärke geschlechtsspezifischer Abweichungen nimmt offenbar mit Voranschreiten der Schullaufbahn zu, so dass z. B. in der Sekundarstufe II größere Leistungsdispositionen zu beobachten sind als in der Sekundarstufe I. Das Vorhandensein solcher Effekte kann auch die vorliegende Arbeit am Beispiel der FALKE-Erhebung für den Bereich der Funktionenlehre und frühen Analysis zu Beginn der Oberstufe bestätigen.

Auf globaler Testebene ergibt sich ein geschlechtsspezifischer Effekt d zu Gunsten der männlichen Probanden dabei einheitlich für beide Tests im Bereich zwischen 0.37 und 0.38. Effekte dieser Größenordnung werden von Hyde (2005) als „mittel“ bewertet, wenngleich diese am unteren Rand des entsprechenden Intervalls von 0.35 bis 0.65 liegen (s. Abschnitt 2.1).

Vergleicht man die einzelnen Schulformen, zeigt sich das Gymnasium aufgrund seines überproportionalen Anteils an der Gesamtstichprobe repräsentativ für diese. So ergeben sich je nach betrachtetem Test und betrachteter Skala ähnliche Werte für d wie für die Gesamtstichprobe zwischen 0.38 und 0.40. Für die anderen beiden Schulformen Gesamtschule und Berufskolleg ergeben sich entgegen der Literatur höhere Werte zwischen 0.53 und 0.64 (s. Abschnitt 2.1 und 2.2). Allein für die Durchführung des zweiten Tests am Berufskolleg zeigen sich die Leistungsdispositionen zwischen den Geschlechtern mit Werten für d zwischen 0.36 und 0.37 etwas geringer als für das Gymnasium, was jedoch auch mit der überproportional hohen Absprungquote zwischen den Tests für diese Schulform zusammenhängen kann. Insgesamt deutet sich aber eine erhöhte Divergenz beider Geschlechter mit zunehmendem Lebensalter an diesen Schulformen an. Bei einer Betrachtung der Effekte

auf Ebene des Gesamttests läuft man Gefahr, diese aufgrund des auftretenden Simpson-Paradoxons zu missachten.

Hierbei können die größeren geschlechtsspezifischen Effekte an Gesamtschulen und Berufskollegs auch Ausdruck der generellen Divergenz während der Schullaufbahn sein, da Schülerinnen und Schüler an Gesamtschulen und Berufskollegs während der Absolvierung der Einführungsphase im Mittel ein Jahr älter sein dürften als entsprechende Gymnasiasten. Sie können aber auch durch strukturelle Eigenschaften, die mit der entsprechenden Schulform einhergehen, aber hier nicht weiter beleuchtet werden können, bedingt sein. In diesem Fall stünden die entsprechenden Ergebnisse im Widerspruch zur betrachteten Literatur, welche größere geschlechtsspezifische Effekte zu Gunsten männlicher Schüler eher an Gymnasien verortet (Abschnitt 2.1 und 2.2).

In der anschließenden Untersuchung auf Ebene der Einzelitems zeigen vor allem Items, die qualitative Funktionen umfassen (insbesondere „Kegelfüllung“, und „Schwimmbecken“), einen großen Abstand der durchschnittlichen Lösungsquoten beider Geschlechter. Am stärksten ist Item „Kegelfüllung“ (J9SD) betroffen, für welches die Lösungsquote für männliche Probanden 19.7 Prozentpunkte höher ausfällt als für weibliche Probanden. Auch das Item „Weihnachtsmann“ (H7ZD) zeigt sich auffällig. Zwar liegen hier beide Geschlechter hinsichtlich der Lösungsquotendifferenz nicht sehr weit auseinander, jedoch ist dieser Wert aufgrund der generell geringen Lösungsquote für dieses Item nicht mit Item J9SD vergleichbar. Für die Aufgabe „Weihnachtsmann“ zeigt sich mit einem Wert von RR = 5.68 jedoch das höchste relative Risiko.

Der potenzielle Lösungserfolg in den beiden genannten Aufgaben dürfte jeweils durch eine ausgeprägte Vorstellungskraft begünstigt werden. So müssen einerseits bewegte Prozesse in Graphen gefasst, andererseits Aussagen hinsichtlich einer proportionalen Flächenvergrößerung getroffen werden. Möglicherweise bilden somit u. a. unterschiedliche Raumvorstellungsfähigkeiten hier einen literaturkonformen Erklärungsansatz der beobachteten Dispositionen (s. Abschnitt 2.4). Da insgesamt jedoch jedes Item des ersten Tests vorteilhaft für das männliche Geschlecht ausfällt und der festgestellte Effekt jeweils nur hinsichtlich seiner Stärke variiert, kann es sich hierbei nur um einen Teilerklärungsansatz handeln, so dass die genaue Ursache der festgestellten geschlechtsspezifischen Differenzen sich nicht vollends aufklären lässt.

Im Gegensatz zum ersten sind für den zweiten Test für wenige Items leichte Leistungsvorsprünge der Schülerinnen zu beobachten. Diese erreichen jedoch nur im Fall von Item „Parabelöffnung“ (D6LG1) übliches Signifikanzniveau. Insgesamt unterscheiden sich die Lösungsquoten hinsichtlich der Geschlechter vor allem für jene Items wenig, welche leicht an erlernte Kalküle auslagerbar scheinen. So weisen neben dem Item „Parabelöffnung“ (D6LG1) vor allem auch Items zum symbolischen Differenzieren geringe geschlechtsspezifische Differenzen auf. In geringem Ausmaß betroffen sind zudem Items zum graphischen Ableiten (U3PT, V3RK, S3AB). Wie bereits in Abschnitt 2.4 geschildert, wird ein besseres oder gleich gutes Abschneiden der Mädchen häufig in solchen Aufgaben beobachtet, für welche ein festes Lösungsprozedere existiert und angewendet werden kann, d.h. für welche erlernbare Kalküle im obigen Sinne existieren (Hyde et al. 1990). Dies bildet somit möglicherweise auch ein Indiz dafür, dass graphisches Differenzieren vor allem in Form von „Rezepten“ vorgenommen wird und somit nicht notwendigerweise ein umfangreich verinnerlichtes Konzept ausgebildet sein muss. Der deutlich spekulative Charakter dieser Überlegung darf dabei jedoch nicht außer Acht gelassen werden.

Die vorliegende Studie liefert somit Einblicke in die sich i. d. R. bei mathematischen Leistungsstudien einstellenden geschlechtsspezifischen Effekte. Durch die Operationalisierung des Instruments nahe an einem mathematischen Inhalt – der Funktionenlehre der Sekundarstufe I sowie der frühen Analysis der Oberstufe und somit dem Funktionalen Denken – lassen sich einige Beobachtungen und daraus abgeleitete Vermutungen zur Bedeutung des Geschlechts in diesem Bereich aufstellen. Diese erstrecken sich insbesondere auf einzelne Aufgabentypen, etwa solche die Füllgraphen thematisieren.

Gerade durch die begrenzte Anzahl an administrierten Items lassen sich jedoch nur bedingt Aussagen zur Verallgemeinerbarkeit der hier festgestellten Ergebnisse treffen. An dieser Stelle sind weitere Untersuchungen nötig. Solche können einerseits durch mehr Variation einzelner Item-Merkmale und einer höheren Itemanzahl in Form ähnlicher quantitativer Designs, aber andererseits auch in Form qualitativer Studiendesigns weitere Einblicke in die Wirkungsweisen geschlechtsspezifischer Leistungsdifferenzen im Bereich des Funktionalen Denkens bieten.

Anmerkungen

[1] Der hier vorliegende Artikel stellt z. T. entsprechende Ergebnisse dieser Dissertation in erweiterter Form dar. Einige Passagen und Abbildungen finden sich daher in ähnlicher oder sogar unveränderter Form auch in dieser Publikation. Auf eine entsprechende Einzelausweisung wird aus Gründen der Lesbarkeit verzichtet.

Reprinted/adapted by permission from Springer: Springer Spektrum, Funktionales Denken beim Übergang von der Funktionenlehre zur Analysis by Klinger, M. (2018)

[2] In der entsprechenden Jahrgangsstufe haben Schülerinnen und Schüler in etwa ein Alter zwischen 16 und 18 Jahren. Obwohl somit die Probandinnen und Probanden kurz vor Abschluss der Adoleszenz stehen, werden im vorliegenden Beitrag die Bezeichner „Mädchen“ bzw. „Junge“ gegenüber den Begriffen „Frau“ bzw. „Mann“ bevorzugt.

[3] Hierbei wurde auf die numerisch-tabellarische Form insgesamt sowie auf die Implementation der Richtung der Darstellungswechsel u.a. zu Gunsten einer Komplexitätsreduktion hinsichtlich der Testentwicklung verzichtet (vgl. Klinger, 2018, S. 184 ff.).

[4] In der vorliegenden Arbeit werden die Begriffe „Berufliches Gymnasium“ und „Berufskolleg“ synonym gebraucht. Gemeint ist stets das Berufliche Gymnasium als Bestandteil eines Berufskollegs.

[5] Eine ausführliche Beschreibung aller Items kann Klinger (2018) entnommen werden. An dieser Stelle sollen nur jene für die Analyse besonders relevanten Aufgaben dargestellt werden. Die verwendeten Testhefte der FALKE-Erhebung sind zudem in Klinger (2017) zu finden.

Literatur

Adams, R. & Carstensen, C. (2002). Scaling outcomes. In R. Adams & M. Wu (Hrsg.), *PISA 2000 technical report* (Kap. 13, S. 149–162). Paris: OECD.

Baumert, J. & Lehmann, R. (1997). *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich: Deskriptive Befunde*. Opladen: Leske+Budrich.

Blum, W. & Törner, G. (1983). *Didaktik der Analysis*. Göttingen: Vandenhoeck & Ruprecht.

Böhme, K. & Roppelt, A. (2012). Geschlechtsbezogene Disparitäten. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011 (S. 170–189). Münster: Waxmann.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. (Hrsg.). (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Brunner, M., Krauss, S. & Martignon, L. (2011). Eine alternative Modellierung von Geschlechtsunterschieden in Mathematik. *Journal für Mathematik-Didaktik, 32*(2), 179–204.

Büchter, A. (2004). Die Wissenschaft hat festgestellt ...! Wie man sich gegen Fehlschlüsse (nicht nur) in der Bildungsforschung wappnet. In G. Eikenbusch & T. Leuders (Hrsg.), *Lehrer-Kursbuch Statistik: Alles über Daten und Zahlen im Schulalltag* (Kap. 6, S. 103–107). Berlin: Cornelsen Scriptor.

Büchter, A. (2008). Funktionale Zusammenhänge erkunden. *mathematik lehren, 148*, 4–10.

Büchter, A. (2010). *Zur Erforschung von Mathematikleistung: Theoretische Studie und empirische Untersuchung des Einflussfaktors Raumvorstellung* (Dissertation, Technische Universität Dortmund, Dortmund).

Burnett, S. A., Lane, D. M. & Dratt, L. M. (1979). Spatial visualization and sex differences in quantitative ability. *Intelligence, 3*(4), 345–354.

Caplan, J. B. & Caplan, P. J. (2005). The perseverative search for sex differences in mathematics ability. In A. M. Gallagher and J. C. Kaufman (Hrsg.), *Gender differences in mathematics: An integrative psychological approach* (Kap. 2, S. 25–47). Cambridge: Cambridge University Press.

Cheema, J. R. & Galluzzo, G. (2013). Analyzing the gender gap in math achievement: Evidence from a large-scale US sample. *Research in Education, 90*(1), 98–112.

Chipman, S. F., Krantz, D. H. & Silver, R. (1992). Mathematics anxiety and science careers among able college women. *Psychological Science, 3*(5), 292–295.

Chipman, S. F., Marshall, S. P. & Scott, P. A. (1991). Content effects on word problem performance: A possible source of test bias? *American Educational Research Journal, 28*(4), 897–915.

Contini, D., Di Tommaso, M. L. & Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review, 58*, 32–42.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale: Lawrence Erlbaum.

Crow, T. J. (1994). The case for an X-Y homologous determinant of cerebral asymmetry. *Cytogenetics and Cell Genetics, 67*(4), 393–394.

Danckwerts, R. & Vogel, D. (2006). *Analysis verständlich unterrichten: Mathematik Primar- und Sekundarstufe*. Heidelberg: Spektrum.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*(3), 279–299.

De Bock, D., Van Dooren, W., Janssens, D. & Verschaffel, L. (2002). Improper use of linear reasoning: An in-depth study of the nature and the irresistibility of secondary school students' errors. *Educational Studies in Mathematics, 50*(3), 311–334.

De Bock, D., Van Dooren, W., Janssens, D. & Verschaffel, L. (2007). *The illusion of linearity: From analysis to improvement*. New York: Springer.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Wainer (Hrsg.), *Differential item functioning* (Kap. 3, S. 35–66). Hillsdale: Lawrence Erlbaum.

Duval, R. (2006). A cognitive analysis of problems of comprehension in a learning of mathematics. *Educational Studies in Mathematics, 61*(1–2), 103–131.

Eccles, J. S., Jacobs, J. E. & Harold, R. D. (1990). Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. *Journal of Social Issues, 46*(2), 183–201.

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.

Ellison, G. & Swanson, A. (2010). The gender gap in secondary school: Mathematics at high achievement levels: Evidence from the American Mathematics Competitions. *Journal of Economic Perspectives, 24*(2), 109–128.

Else-Quest, N. M., Hyde, J. S. & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics:

A meta-analysis. *Psychological Bulletin, 136*(1), 103–127.

Else-Quest, N. M., Mineo, C. C., Higgins, A. (2013). Math and science attitudes and achievement at the intersection of gender and ethnicity. *Psychology of Women Quarterly, 37*(3), 293–309.

Embretson, S. E. & Reise, S. P. (2009). *Item response theory for psychologists.* New York: Psychology Press.

Fennema, E. (1974). Mathematics learning and the sexes: A review. *Journal for Research in Mathematics Education, 5*(3), 126–139.

Fennema, E. & Peterson, P. L. (1987). Effective teaching for boys and girls. In D. C. Berliner & B. V. Rosenshine (Hrsg.), *Talks to teachers: A Festschrift for N. L. Gage* (S. 111–125). New York: Random House.

Fennema, E., Peterson, P. L., Carpenter, T. P. & Lubinski, C. A. (1990). Teachers' attributions and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics, 21*(1), 55–69.

Fennema, E. H. & Sherman, J. A. (1978). Sex-related differences in mathematics achievement and related factors: A further study. *Journal for Research in Mathematics Education, 9*(3), 189–203.

Fox, L. H., Fennema, E. & Sherman, J. (1977) (Hrsg.). *Women in mathematics: Research perspectives for change*. Washington, D.C.: National Institute of Education.

Fryer, R. G. & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal, 2*(2), 210–240.

Ganley, C. M. & Vasilyeva, M. (2014). The role of anxiety and working memory in gender differences in mathematics. *Journal of Educational Psychology, 106*(1), 105–120.

Geary, D. C. (1989). A model for representing gender differences in the pattern of cognitive abilities. *American Psychologist, 44*(8), 1155–1156.

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences, 19*(2), 229–284.

Good, C., Aronson, J. & Harder, J. A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology, 29*(1), 17–28.

Greefrath, G., Oldenburg, R., Siller, H.-S., Ulm, V. & Weigand, H.-G. (2016a). *Didaktik der Analysis: Aspekte und Grundvorstellungen zentraler Begriffe*. Berlin: Springer Spektrum.

Greefrath, G., Oldenburg, R., Siller, H.-S., Ulm, V. & Weigand, H.-G. (2016b). Aspects and "Grundvorstellungen" of the concepts of derivative and integral: Subject matter-related didactical perspectives of concept formation. *Journal für Mathematik-Didaktik, 37*(1), 99–129.

Grüßing, M. (2012). *Räumliche Fähigkeiten und Mathematikleistung Räumliche Fähigkeiten und Mathematikleistung: Eine empirische Studie mit Kindern im 4. Schuljahr*. Münster: Waxmann.

Guiso, L., Monte, F., Sapienza, P. & Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880), 1164–1165.

Gunderson, E. A., Ramirez, G., Levine, S. C. & Beilock, S. L. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles, 66*(3–4), 153–166.

Hahn, S. (2008). *Bestand und Änderung: Grundlegung einer vorstellungsorientierten Differentialrechnung*. Oldenburg: Didaktisches Zentrum.

Hahn, S. & Prediger, S. (2008). Bestand und Änderung – Ein Beitrag zur Didaktischen Rekonstruktion der Analysis. *Journal für Mathematik-Didaktik, 29*(3–4), 163–198.

Halpern, D. F., Wai, J. & Saw, A. (2005). A psychobiosocial model: Why females are sometimes greater than and sometines less than males in math achievement. In A. M. Gallagher and J. C. Kaufman (Hrsg.), *Gender differences in mathematics: An integrative psychological approach* (Kap. 3, S. 48–72). Cambridge: Cambridge University Press.

Halpern, D. F. & Wright, T. M. (1996). A process-oriented model of cognitive sex differences. *Learning and Individual Differences, 8*(1), 3–24.

Harris, A. M. & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*(2), 137–151.

Hedges, L. V. & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41–45.

Hußmann, S. & Prediger, S. (2010). Vorstellungsorientierte Analysis – auch in Klassenarbeiten und zentralen Prüfungen. *Praxis der Mathematik in der Schule, 52*(31), 35–38.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581–592.

Hyde, J. S., Fennema, E. & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139–155.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B. & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321*(5888), 494–495.

Ivanov, S. (2011). Mathematische Kompetenz und Einstellungen zum Mathematikunterricht. In U. Vieluf, S. Ivanov & R. Nikolova (Hrsg.), *KESS 10/11: Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der Sekundarstufe I und zu Beginn der gymnasialen Oberstufe* (Kap. 4, S. 75–120). Münster: Waxmann.

Janvier, C. (1978). *The interpretation of complex cartesian graphs representing situations: Studies and teaching experiments* (Dissertation, University of Nottingham, Nottingham).

Klieme, E. (1986). Zur Problematik geschlechtsspezifischer Mathematikleistungen. In H.-G. Steiner (Hrsg.), *Grundfragen der Entwicklung mathematischer Fähigkeiten* (S. 133–151). Köln: Aulis Deubner.

Klinger, M. (2017). *Funktionales Denken beim Übergang von der Funktionenlehre zur Analysis* (Anhang zur Dissertation). Abgerufen von http://doi.org/10.17185/duepublico/42711 (letzter Zugriff: 23.02.2019).

Klinger, M. (2018). *Funktionales Denken beim Übergang von der Funktionenlehre zur Analysis: Entwicklung eines Testinstruments und empirische Befunde aus der gymnasialen Oberstufe*. Wiesbaden: Springer Spektrum.

Klinger, M. & Barzel, B. (2018a). Zum Einfluss des Geschlechts beim Darstellungswechsel funktionaler Zusammenhänge. In Fachgruppe Didaktik der Mathematik der Universität Paderborn (Hrsg.), *Beiträge zum Mathematikunterricht 2018* (Bd. 2, S. 987–990). Münster: WTM-Verlag.

Klinger, M. & Barzel, B. (2018b). Zielgerichtete Entwicklung von verstehensorientierten Leistungstestaufgaben am Beispiel des Funktionalen Denkens in der frühen Analysis der Oberstufe. In Fachgruppe Didaktik der Mathematik der Universität Paderborn (Hrsg.), *Beiträge zum Mathematikunterricht 2018* (Bd. 2, 983–986). Münster: WTM-Verlag.

Klinger, M., Thurm, D., Barzel, B., Greefrath, G. & Büchter, A. (2018). Lehren und Lernen mit digitalen Werkzeugen: Entwicklung und Durchführung einer Fortbildungsreihe. In R. Biehler, T. Lange, T. Leuders, P. Scherer, B. Rösken-Winter & C. Selter (Hrsg.), *Mathematikfortbildungen professionalisieren: Konzepte, Beispiele und Erfahrungen des Deutschen Zentrums für Lehrerbildung Mathematik* (Kap. 20, S. 395–416). Wiesbaden: Springer Spektrum.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland) (2015). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife* (Beschluss der Kultusministerkonferenz vom 18.10.2012). Köln: Kluwer.

Köller, O. & Klieme, E. (2000). Geschlechtsdifferenzen in den mathematisch-naturwissenschaftlichen Leistungen. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn / Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (Kap. 9, Bd. 2, S. 373–404). Opladen: Leske+Budrich.

Krabbendam, H. (1982). The non-quantitative way of describing relations and the role of graphs: Some experiments. In G. van Barneveld & H. Krabbendam (Hrsg.), *Conference on functions: Report 1* (S. 125–146). Enschede: Foundation for Curriculum Development.

Laakmann, H. (2013). *Darstellungen und Darstellungswechsel als Mittel zur Begriffsbildung: Eine Untersuchung in rechner-unterstützten Lernumgebungen*. Wiesbaden: Springer Spektrum.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test language for PISA science items. *International Journal of Testing, 9*(2), 122–133.

Leder, G. C. & Forgasz, H. J. (2018). Measuring who counts: Gender and mathematics assessment. *ZDM Mathematics Education, 50*(4), 687–697.

Leuders, T. & Prediger, S. (2005). Funktioniert's? – Denken in Funktionen. *Praxis der Mathematik in der Schule, 47*(2), 1–7.

Linacre, J. M. (2002). Differential item functioning and differential test functioning (DIF & DTF). *Rasch Measurement Transactions, 16*(3), 889.

Lindberg, S. M., Hyde, J. S., Petersen, J. L. & Linn, M. C. (2010). *New trends in gender and mathematics performance: A meta-analysis. Psychological Bulletin, 136*(6), 1123–1135.

Liu, L., Wilson, M. & Paek, I. (2008). A multidimensional Rasch analysis of gender differences in PISA mathematics. *Journal of Applied Measurement, 9*(1), 18–35.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah: Lawrence Erlbaum.

Malle, G. (2000). Zwei Aspekte von Funktionen: Zuordnung und Kovariation. *mathematik lehren, 103*, 8–11.

Marsh, H. W. & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal, 35*(4), 705–738.

Mullis, I. V. S., Martin, M. O., Foy, P. & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill: Boston College.

Muzzatti, B. & Agnoli, F. (2007). Gender and mathematics: Attitudes and stereotype threat susceptibility in Italian children. *Developmental Psychology, 43*(3), 747–759.

Neuburger, S., Jansen, P., Heil, M. & Quaiser-Pohl, C. (2012). A threat in the classroom: Gender stereotype activation and mental-rotation performance in elementary-school children. *Zeitschrift für Psychologie, 220*(2), 61–69.

Nitsch, R. (2015). *Diagnose von Lernschwierigkeiten im Bereich funktionaler Zusammenhänge: Eine Studie zu typischen Fehlermustern bei Darstellungswechseln*. Wiesbaden: Springer Spektrum.

OECD (Organisation for Economic Co-operation and Development). (2014). *PISA 2012 Ergebnisse: Was Schülerinnen und Schüler wissen und können*. München: Bertelsmann.

Pajares, F. (1996). Self-efficacy beliefs and mathematical problem-solving of gifted students. *Contemporary Educational Psychology, 21*(4), 325–344.

Pant, H. A., Stanat, P., Pöhlmann, C. & Böhme, K. (2013). Die Bildungsstandards im allgemeinbildenden Schulsystem. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (Kap. 1, S. 2–21). Münster: Waxmann.

Penner, A. M. (2003). International gender × item difficulty interactions in mathematics and science achievement tests. *Journal of Educational Psychology, 95*(3), 650–655.

Penner, A. M. & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research, 37*(1), 239–253.

Quinn, D. M. & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57*(1), 55–71.

Raju, N. S., van der Linden, W. J. & Fleer, P. F. (1995). IRT-based internal measures of differential functioning items and tests. *Applied Psychological Measurement, 19*(4), 353–368.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reilly, D., Neumann, D. L. & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology, 107*(3), 645–662.

Robinson, J. P. & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*(2), 268–302.

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M. & Copur-Gencturk, Y. (2014). Teachers' perceptions of

students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology, 50*(4), 1262–1281.

Ryan, K. E. & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*(1), 73–90.

Schlöglhofer, F. (2000). Vom Foto-Graph zum Funktions-Graph. mathematik lehren, Heft 103, 16–17.

Schroeders, U., Penk, C., Jansen, M. & Pant, H. A. (2013). Geschlechtsbezogene Disparitäten. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (Kap. 7, S. 249–274). Münster: Waxmann.

Schwery, D., Hulac, D. & Schweinle, A. (2016). Understanding the gender gap in mathematics achievement: The role of self-efficacy and stereotype threat. School *Psychology Forum, 10*(4), 386–396.

Sherman, J. A. (1967). Problem of sex differences in space perception and aspects of intellectual functioning. *Psychological Review, 74*(4), 290– 299.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B, 13*(2), 238–241.

Spencer, S. J., Steele, C. M. & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4–28.

Stellmacher, H. (1986). Die nichtquantitative Beschreibung von Funktionen durch Graphen beim Einführungsunterricht. In G. von Harten, H. N. Jahnke, T. Mormann, M. Otte, F. Seeger, H. Steinbring & H. Stellmacher (Hrsg.), *Funktionsbegriff und funktionales Denken* (Kap. 2, S. 21–34). Köln: Aulis Deubner.

Stewart, C., Root, M. M., Koriakin, T., Choi, D., Luria, S. R., Bray, M. A., Sassu, K., Maykel, C., O'Rourke, P. & Courville, T. (2017). Biological gender differences in students' errors on mathematics achievement tests. *Journal of Psychoeducational Assessment, 35*(1–2), 47–56.

Stoet, G. & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology, 16*(1), 93–103.

Swan, M. (1982). The teaching of functions and graphs. In G. van Barneveld & H. Krabbendam (Hrsg.), *Conference on functions: Report 1* (S. 151–165). Enschede: Foundation for Curriculum Development.

Swan, M. (Hrsg.). (1985). *The language of functions and graphs: An examination module for secondary schools*. Nottingham: Shell Centre for Mathematical Education.

Tartre, L. A. & Fennema, E. (1995). Mathematics achievement and gender: A longitudinal study of selected cognitive and affective variables [grades 6–12]. *Educational Studies in Mathematics, 28*(3), 199–217.

Thurm, D., Klinger, M. & Barzel, B. (2015). How to professionalize teachers to use technology in a meaningful way – Design research of a CPD program. In N. Amado & S. Carreira (Hrsg.), *Proceedings of the 12th International Conference on Technology in Mathematics Teaching* (S. 335–343). Faro: Universidade do Algarve.

vom Hofe, R., Lotz, J. & Salle, A. (2015). Analysis: Leitidee Zuordnung und Veränderung. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (Kap. 6, S. 149–184). Berlin: Springer Spektrum.

Vollrath, H.-J. (1989). Funktionales Denken. *Journal für Mathematik-Didaktik, 10*(1), 3–37.

Voyer, D. (1996). The relation between mathematical achievement and gender differences in spatial abilities: A suppression effect. *Journal of Educational Psychology, 88*(3), 563–571.

Voyer, D. (1998). Mathematics, gender, spatial performance, and cerebral organization: A suppression effect in talented students. *Roeper Review, 20*(4), 251–258.

Walsh, M., Hickey, C. & Duffy, J. (1999). Influence of item content and stereotype situation on gender differences in mathematical problem solving. *Sex Roles, 41*(314), 219–240.

Wigfield, A. & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review, 12*(3), 265–310.

Xie, Y. & Shauman, K. A. (2003). *Women in science: Career processes and outcomes*. Cambridge: Harvard University Press.

Zohar, A. & Gershikov, A. (2008). Gender and performance in mathematical tasks: Does the context make a difference? *International Journal of Science and Mathematics Education, 6*(4), 677–693.

Anschrift des Verfassers

Dr. Marcel Klinger
Universität Duisburg-Essen
Fakultät für Mathematik
Thea-Leymann-Str. 9
45127 Essen
marcel.klinger@uni-due.de