Die Rolle kognitiver Prozesse beim komplexen Problemlösen

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt beim Fachbereich 05

Psychologie und Sportwissenschaften

der Johann Wolfgang Goethe-Universität

Frankfurt am Main

von

Beate Eichmann

aus Bad Pyrmont

Frankfurt am Main 2019

(D 30)

vom Fachbereich 05 der

Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekanin: Prof. Dr. Sonja Rohrmann

Gutachter: Prof. Dr. Frank Goldhammer

Prof. Dr. Holger Horz

Datum der Disputation: 16.09.2020

INHALTSVERZEICHNIS

ZUSAMMENFASSUNG

Das Ziel der vorliegenden Arbeit ist die Identifikation von leistungsrelevanten kognitiven Prozessen beim komplexen Problemlösen (KPL). Außerdem soll untersucht werden, ob sich Leistungsunterschiede beim KPL zwischen soziodemografischen Gruppen durch Prozessmaße erklären lassen. Dazu wurden in den drei Einzelarbeiten, auf denen diese Arbeit basiert, verschiedene Prozesse und ihr Zusammenhang mit der Leistung beim KPL untersucht. Darüber hinaus schafft die vorliegende Arbeit einen theoretischen Rahmen, in den sich die drei Einzelarbeiten einordnen lassen.

Die Fähigkeit komplexe Probleme lösen zu können, ist eine grundlegende Kompetenz in Bildung und Alltag und ermöglicht eine aktive Teilhabe an der Gesellschaft. KPL kann daher auch als Schlüsselkompetenz in der Wissensgesellschaft des 21. Jahrhunderts verstanden werden (Binkley et al., 2012; Trilling & Fadel, 2009). Komplexe Probleme begegnen jedem Menschen im beruflichen und privaten Umfeld sowie auf gesellschaftlicher Ebene. Daher ist es wichtig zu verstehen, welche Prozesse für effektives KPL relevant sind. Darüber hinaus wurden wiederholt Leistungsunterschiede beim KPL in Abhängigkeit vom Geschlecht und vom Migrationshintergrund der Personen festgestellt (OECD, 2014a; Sonnleitner, Brunner, Keller & Martin, 2014; Wüstenberg, Greiff, Molnár & Funke, 2014).

In der ersten Arbeit wird der Zusammenhang verschiedener Aspekte von Planung mit der Leistung beim KPL untersucht. Die betrachteten Planungsaspekte sind die Dauer des längsten Planungsintervalls, der Zeitpunkt zu dem Planung erfolgt und die Variation der Dauer von Planungsintervallen im Problemlöseprozess. Zudem wird untersucht, ob die Effekte bei verschiedenen Aufgaben unterschiedlich ausgeprägt sind und ob es Interaktionseffekte der drei Planungsaspekte gibt. Die Ergebnisse zeigen, dass Planung grundsätzlich zu einem möglichst frühen Zeitpunkt stattfinden sollte. Die beiden anderen Planungsaspekte wiesen hingegen aufgabenabhängige Effekte auf. Außerdem gab es Interaktionseffekte. Insgesamt wurde bei leichten KPL-Aufgaben festgestellt, dass ähnlich wie beim analytischen Problemlösen Planung zu einem frühen Zeitpunkt einen positiven Einfluss auf die Leistung hat (Unterrainer & Owen, 2006). Auch der Einfluss der Variation der Planungsdauer hing mit der Aufgabenschwierigkeit zusammen, wobei bei leichten Aufgaben ein gleichmäßiges und bei schweren Aufgaben ein ungleichmäßigeres Vorgehen vorteilhaft war. Der Effekt der Planungsdauer war ebenfalls aufgabenabhängig, jedoch nur schwach mit der Aufgabenschwierigkeit korreliert. Somit scheinen andere Aufgabeneigenschaften für diesen Zusammenhang ursächlich zu sein.

In der zweiten Arbeit werden Leistungsunterschiede beim KPL in Abhängigkeit vom Geschlecht und vom Migrationshintergrund der Schülerinnen und Schüler untersucht. Das Ziel dieser Arbeit ist es, Leistungsunterschiede zwischen diesen Gruppen durch Prozessmaße zu erklären. Da es Evidenz für einen Zusammenhang der Häufigkeit von Interaktion beziehungsweise Exploration mit der Leistung beim KPL gibt, werden diese als Prozessmaße verwendet (Bell & Kozlowski, 2008; Dormann & Frese, 1994; Naumann, Goldhammer, Rölke & Stelter, 2014). Erwartungskonform wurden Leistungsunterschiede beim KPL zugunsten von Jungen gegenüber Mädchen und zugunsten von Schülerinnen

und Schülern ohne Migrationshintergrund gegenüber Schülerinnen und Schülern mit Migrationshintergrund festgestellt. Außerdem zeigte sich, dass beide Prozessmaße positiv mit der KPL-Leistung korrelierten. Der Leistungsunterschied zwischen Jungen und Mädchen konnte durch die Interaktionshäufigkeit teilweise und durch die Explorationshäufigkeit vollständig aufgeklärt werden. Der Leistungsunterschied in Abhängigkeit des Migrationshintergrundes konnte hingegen durch keines der beiden Maße erklärt werden.

Die dritte Arbeit hat zum einen das Ziel, die Rolle von Explorationsverhalten beim KPL genauer zu klären. Zum anderen werden mit einem explorativen Ansatz komplexe Verhaltensmuster untersucht. Dazu wurde eine weitere Differenzierung von Exploration in lösungsrelevante und lösungsunabhängige Exploration vorgenommen. Es konnte gezeigt werden, dass im Gegensatz zu den Ergebnissen aus der zweiten Arbeit lösungsunabhängige Exploration vermehrt bei erfolgloser Aufgabenbearbeitung auftritt. Lediglich lösungsrelevante Exploration scheint also zu einer höheren KPL-Leistung beizutragen. Zudem wurden verschiedene Verhaltensmuster identifiziert, die auf konkrete Stärken und Schwächen im komplexen Problemlöseprozess von Schülerinnen und Schülern hinweisen.

Die vorliegende Arbeit erweitert die theoretische Basis für KPL, indem sie kognitive Prozesse ordnet und im Sinne einer Intention interpretierbar macht. Weiterhin werden durch die empirischen Arbeiten Erkenntnisse über die Relevanz der untersuchten Prozesse für die Leistung beim KPL und für die Erklärung von Leistungsunterschieden gewonnen. Damit erleichtert diese Arbeit die Erklärung der Rolle kognitiver Prozesse beim KPL, um so das Verständnis dieses Konstruktes zu verbessern. Dies ist wiederum die Basis, um Schülerinnen und Schüler beim Erwerb der Kompetenz zum Lösen komplexer Probleme zu unterstützen und sie so auf die Herausforderungen des 21. Jahrhunderts vorzubereiten.

1. EINLEITUNG

„Es ist eine fast schon triviale Floskel, wenn man sagt, daß wir in einer Welt leben, die immer komplexer wird.“ Dies schrieben Dörner, Kreuzig, Reither und Stäudel bereits 1983. Die Autoren bezogen sich mit dieser Aussage vor allem auf Prozesse der Globalisierung wie die Abhängigkeit der europäischen Energieversorgung und Konjunktur von der politischen Lage im Nahen Osten. Die von Dörner et al. (1983) angeführten Beispiele sind jedoch auch heute noch aktuell. Als Gesellschaft sehen wir uns immer häufiger komplexen Problemen gegenüber. Ein Beispiel für ein globales, komplexes Problem ist der aktuelle Klimawandel. Dieses Problem umfasst eine große Anzahl von Einflussgrößen, die miteinander in Verbindung stehen (z.B. die Fläche von Permafrostböden, der CO_2-Gehalt der Atmosphäre und die globale Durchschnittstemperatur). Die Grundlage, um ein komplexes Problem wie dieses zu lösen, ist die Zusammenhänge zwischen den beteiligten Variablen zu verstehen, um in der Lage zu sein, das System in die gewünschte Richtung zu lenken. Bei einem System wie dem Weltklima sind viele Zusammenhänge jedoch nicht vollständig bekannt. Komplexes Problemlösen (KPL) bedeutet, die (teils) unbekannten Zusammenhänge von Ursachen und Wirkungen zu ergründen und zu nutzen, um die gewünschten Ziele zu erreichen. Diese Fähigkeit ist notwendig, sowohl um globale Probleme zu lösen als auch um in einer sich ständig verändernden Welt bestehen zu können (Koeppen, Hartig, Klieme & Leutner, 2008). Durch Prozesse wie die wachsende Digitalisierung und die steigende Mobilität der Menschen hat sich die Komplexität nicht nur noch weiter vergrößert; sie hat sich auch von der politischen und wirtschaftlichen Ebene in das Leben jedes Einzelnen ausgebreitet (Bos, Eickelmann & Gerick, 2014; Trilling & Fadel, 2009). Beispiele für komplexe Alltagsprobleme können das Orientieren in einer unbekannten Stadt, das Bedienen technischer Geräte oder die Auswahl des richtigen Studiengangs sein. Sowohl privat als auch im Berufsleben sind wir täglich mit komplexen Problemen konfrontiert. Da wir heute jedoch nicht wissen können, welche Anforderungen in einer zukünftige Welt bestehen werden, ist es gerade für junge Menschen entscheidend, dass sie mit komplexen, unbekannten Problemen umgehen können (Greiff, Holt & Funke, 2013; Trilling & Fadel, 2009). Schon heute wird KPL im Berufsleben immer wichtiger, da einfachere Tätigkeiten zunehmend von Maschinen übernommen werden, während Menschen immer komplexere Tätigkeiten ausführen (Autor, Levy & Murnane, 2003; Frey & Osborne, 2013). Aus diesem Grund gilt KPL als eine Schlüsselkompetenz in der Wissensgesellschaft des 21. Jahrhunderts (Binkley et al., 2012; Leutner, Funke, Klieme & Wirth, 2005; OECD, 2013).

Aufgrund ihrer Relevanz wurde Problemlösekompetenz als Testdomäne in das Programme for International Student Assessment (PISA) der OECD aufgenommen (OECD, 2014a). Laut der OECD (2013) ist Problemlösekompetenz die Grundlage für zukünftiges Lernen, Teilhabe an der Gesellschaft und das Erreichen persönlicher Ziele. Jedoch zeigte PISA 2012, dass in den teilnehmenden OECD-Ländern etwa ein Fünftel der Schülerinnen und Schüler nur sehr einfache Probleme lösen konnten, während es 11,4% der Schülerinnen und Schüler in die Gruppe mit der besten Leistung schafften. Neben diesen allgemeinen Leistungsunterschieden zwischen Schülerinnen und Schülern wurden außerdem Leistungsunterschiede zwischen verschiedenen Gruppen festgestellt. Faktoren, die mit der Problemlöseleistung in PISA im Zusammenhang stehen, sind Geschlecht,

sozioökonomischer Status und Migrationshintergrund. Jungen zeigten im OECD-Schnitt eine höhere Problemlöseleistung als Mädchen, Schülerinnen und Schüler mit niedrigem sozioökonomischen Status sowie Schülerinnen und Schüler mit Migrationshintergrund zeigten jeweils eine geringere Leistung als ihre Mitschülerinnen und Mitschüler mit hohem sozioökonomischen Status bzw. ohne Migrationshintergrund.

Diese Leistungsunterschiede werfen Fragen nach der Gerechtigkeit in den betreffenden Bildungssystemen auf. Vor diesem Hintergrund stellt sich die Frage, wie Leistungsunterschiede beim KPL zustande kommen. Komplexe Probleme zu lösen, erfordert eine aktive Auseinandersetzung der problemlösenden Person mit dem Problem (Frensch & Funke, 1995). Daher scheint eine Untersuchung der beim KPL relevanten Prozesse vielversprechend, um Leistungsunterschiede (sowohl zwischen Individuen als auch zwischen Gruppen) zu erklären. Im Rahmen dieser Arbeit werden für die Leistung beim KPL relevante Prozesse identifiziert. Dazu soll analysiert werden, wie sich die kognitiven Prozesse von mehr und weniger erfolgreichen Problemlösenden voneinander unterscheiden. Anschließend wird geprüft, ob diese Prozesse auch Leistungsunterschiede zwischen soziodemografischen Gruppen erklären können. Ziel dieser Arbeit ist es also, das Wissen über relevante Prozesse beim KPL zu erweitern. Dadurch sollen die Grundlagen geschaffen werden, um beim KPL weniger erfolgreiche Personen bei der Verbesserung ihrer Leistung zu unterstützen. Die vorliegende Arbeit basiert auf drei Einzelarbeiten (Arbeit 1, Arbeit 2 und Arbeit 3).

In Kapitel 2 wird zunächst das Konstrukt KPL näher beleuchtet, bevor in Kapitel 3 die Forschungsfragen auf der Grundlage bisheriger Forschungsarbeiten hergeleitet werden. In Kapitel 0 werden die verwendeten Methoden und Ergebnisse der drei empirischen Arbeiten vorgestellt, auf denen diese Arbeit basiert. Die Ergebnisse werden in Kapitel 5 übergreifend diskutiert.

2. KOMPLEXES PROBLEMLÖSEN

2.1. DEFINITION UND ABGRENZUNG

Ein Problem liegt nach Mayer und Wittrock (2006) dann vor, wenn ein Lebewesen ein Ziel erreichen möchte, aber nicht weiß, wie es dieses Ziel erreichen kann. Das Problem besteht aus einem aktuellen Zustand, einem Zielzustand und Barrieren, die den Übergang vom aktuellen in den Zielzustand verhindern (Funke, 2010). Anhand dieser Definitionen wird deutlich, dass eine Situation nie aus sich selbst heraus ein Problem darstellt, sondern dass sich das Problem aus dem Zusammenspiel der Situation und einer Person (oder einem Lebewesen) und deren Zielen ergibt. Eine Situation kann also für eine Person ein Problem darstellen, während sie es für eine andere Person nicht tut. Das Modell von Frensch und Funke (1995), das in Abbildung 2-1 dargestellt ist, verdeutlicht die Relevanz von Personen- und Aufgabenmerkmalen und integriert zusätzlich Umweltbedingungen, die sowohl die Person als auch die Aufgabe beeinflussen können. Somit stellt sich Problemlösen als Zusammenspiel von Person, Aufgabe und Umwelt dar. Probleme lassen sich weiterhin in analytische und komplexe Probleme einteilen. Während analytische Probleme allein durch Nachdenken gelöst werden können, erfordern komplexe Probleme, dass die problemlösende Person mit dem Problem interagiert. Frensch und Funke definieren KPL wie folgt:

CPS [komplexes Problemlösen] occurs to overcome barriers between a given state and a desired goal state by means of behavioral and/or cognitive, multistep activities. The given state, goal state, and barriers between given state and goal state are complex, change dynamically during problem solving, and are intransparent. The exact properties of the given state, goal state, and barriers are unknown to the solver at the outset. CPS implies the efficient interaction between a solver and the situational requirements of the task, and involves a solver's cognitive, emotional, personal, and social abilities and knowledge. (Frensch & Funke, 1995, S. 36)

Komplexe Probleme zeichnen sich laut Frensch und Funke (1995) also durch Komplexität (eine große Anzahl zusammenhängender Variablen), Dynamik (die Veränderung des Problems über die Zeit) und Intransparenz (die Unvollständigkeit der Informationen zu Beginn des Problems) aus. Diese Definition weist, wie viele andere Definitionen von KPL, die scheinbare Tautologie auf, dass komplexes Problemlösen als komplex definiert wird. Der Begriff „Komplexität" bezieht sich jedoch auf zwei unterschiedliche Eigenschaften von Problemen:

1. Komplexität ist eine quantifizierbare Problemeigenschaft, die je nach Definition auch bei „nicht-komplexen" Problemen auftreten kann. Unter Komplexität werden häufig die Anzahl der am Problem beteiligten Variablen und/oder deren Zusammenhänge verstanden (Stadler, Niepel & Greiff, 2019). Andere Autoren definieren Komplexität hingegen als die Anzahl gleichzeitig zu verarbeitender Informationseinheiten (Beckmann & Goode, 2017).

2. Probleme werden in Abgrenzung zu analytischen Problemen als komplex charakterisiert. Die „Komplexität" ist in diesem Kontext eine qualitative Eigenschaft der

Probleme, die die dichotome Einteilung in analytische und komplexe Probleme ermöglicht. Nach der oben genannten Definition sind es Dynamik, Intransparenz und Komplexität (im quantitativen Sinne), die komplexe von analytischen Problemen unterscheiden. Welche Eigenschaften KPL von analytischem Problemlösen abgrenzen, darüber herrscht in der

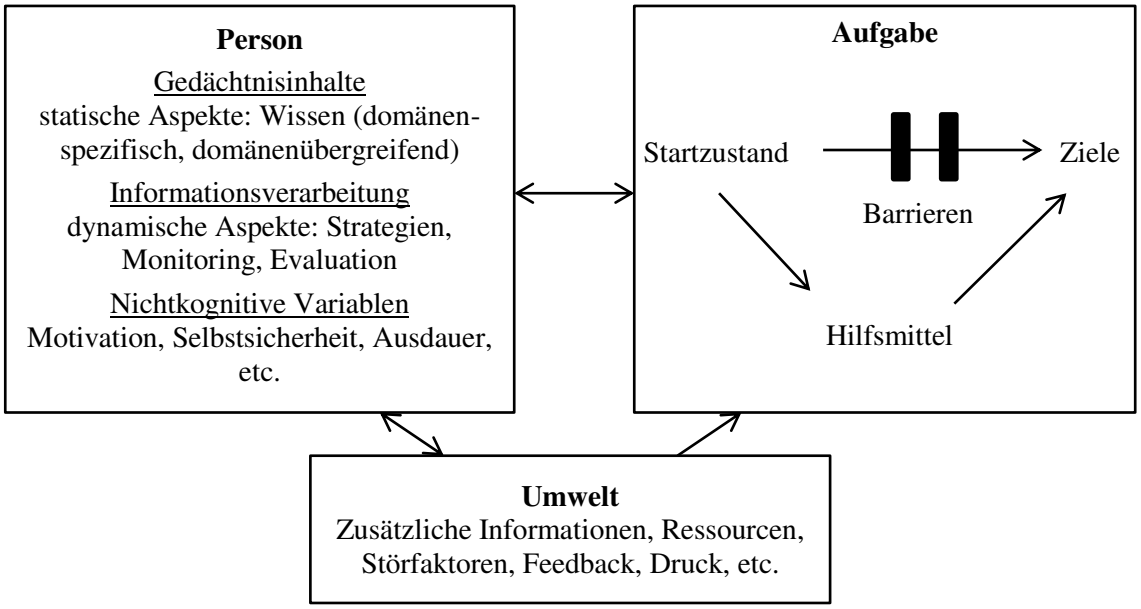

ABBILDUNG 2-1: MODELL EINER KOMPLEXEN PROBLEMLÖSESITUATION NACH FRENSCH UND FUNKE (1995, S. 18)

Literatur jedoch Uneinigkeit. Während Stadler und Niepel et al. (2019) Vernetztheit und Dynamik für die Abgrenzung von KPL und analytischem Problemlösen heranziehen, nennen Fischer und Neubert (2015) die Anzahl der beteiligten Variablen, deren Vernetztheit und die Intransparenz des Problems als zentrale Unterscheidungsmerkmale. Worauf jedoch die meisten Definitionen von KPL inklusive der oben genannten von Frensch & Funke, 1995 hinauslaufen, ist die Notwendigkeit der Interaktion zwischen Person und Problem, die komplexe von analytischen Problemen abgrenzen. Als Interaktion wird in diesem Zusammenhang die Ausführung beobachtbarer Handlungen einer Person mit der Aufgabe verstanden.

Es gibt jedoch auch Kritik am qualitativen Begriff der Komplexität in Abgrenzung zu analytischem Problemlösen. Dörner und Funke (2017) kritisierten die Verwendung des Begriffs „komplex“ für sogenannte minimal komplexe Systeme (siehe Kapitel 2.4.2). Diese werden in der aktuellen KPL-Forschung häufig verwendet (Greiff, Wüstenberg & Avvisati, 2015; Greiff, Niepel, Scherer & Martin, 2016). Die Kritik von Dörner und Funke (2017) richtet sich vor allem darauf, dass in diesen Aufgaben meist nur lineare Zusammenhänge zwischen Variablen simuliert werden. Lineare Zusammenhänge seien laut Dörner und Funke (2017) jedoch nicht ausreichend, um Komplexität adäquat abzubilden. Nach ihrer Definition seien auch der Einsatz von Vorwissen und eine unklare Zielsetzung beim Problemlösen erforderlich, um dieses als komplex zu klassifizieren. Aufgrund dieser Kritik werden für minimal komplexe Systeme teilweise auch Begriffe wie „dynamisch“ statt „komplex“ verwendet (Funke & Greiff, 2017). Allerdings ist der Begriff komplexes Problemlösen (bzw. die englische Entsprechung „complex problem solving“) in der Literatur auch für minimal komplexe Systeme sehr weit verbreitet (Beckmann, Birney

& Goode, 2017; Funke, 2010; Greiff, Kretzschmar, Müller, Spinath & Martin, 2014; Neubert, Kretzschmar, Wüstenberg & Greiff, 2015). Deshalb wird in der vorliegenden Arbeit der Begriff komplexes Problemlösen für das Lösen aller Probleme, die sich nach der oben genannten Definition von Frensch und Funke (1995) klassifizieren lassen, verwendet. Es wird nicht zwischen komplexem und dynamischem Problemlösen unterschieden. In der vorliegenden Arbeit werden daher alle Probleme als komplex verstanden, die aufgrund ihrer Intransparenz oder Dynamik eine Interaktion zwischen Problem und Person erfordern und nicht allein durch Nachdenken gelöst werden können.

2.2. Prozesse des komplexen Problemlösens

Um zu verstehen, welche Prozesse mit der Leistung beim KPL zusammenhängen, werden im Folgenden einige theoretische Ansätze zu Prozessen beim KPL betrachtet. Sowie es keine einheitliche Definition von KPL gibt, so gibt es ebenfalls keine allgemeine Theorie über die Prozesse, die für KPL relevant sind (Beckmann & Goode, 2017; Greiff, 2012a). Verschiedene Autoren nehmen an, dass verschiedene kognitive, metakognitive, motivationale und emotionale Prozesse beim KPL auftreten. Die betrachteten Prozesse sind dabei uneinheitlich und setzen teils auf verschiedenen Ebenen an. Im Folgenden werden die für diese Arbeit wichtigsten Modelle vorgestellt und diskutiert. Anschließend wird aus den bisherigen Modellen ein integriertes Modell der Prozesse beim KPL abgeleitet.

2.2.1. Theoretische Annahmen

Im Folgenden werden einige Modelle vorgestellt, die Prozesse beim KPL aufzeigen. Im Gegensatz zu dem in Kapitel 2.1 vorgestellten Modell von Frensch und Funke (1995), das statische Aspekte der Problemlösesituation beschreibt, geht es in den folgenden Theorien und Modellen um die beim Problemlösen ablaufenden Prozesse. Es geht also darum, wie sich das System aus Person, Aufgabe und Umwelt verändert und wie die Person aktiv in diesen Prozess eingreifen kann, um ihre Ziele zu erreichen.

Eines der frühesten Modelle zu kognitiven Prozessen beim KPL ist das Modell der „Stationen des Planens und Handelns“ beim KPL von Dörner (1989). In diesem Modell stellt Dörner den Problemlöseprozess in mehreren Phasen dar. Dabei folgen die einzelnen Phasen keinem streng linearen Ablauf, sondern es kann laut Dörner zu jedem Zeitpunkt zu einer früheren Phase zurückgekehrt werden. Dörners Modell umfasst die Phasen (1) Zielausarbeitung, (2) Modellbildung und Informationssammlung, (3) Prognose und Extrapolation, (4) Planung von Aktionen, Entscheidung und Durchführung der Aktionen und (5) Effektkontrolle und Revision der Handlungsstrategien.

Die Phase der (1) Zielausarbeitung ergibt sich aus Dörners Verständnis von KPL, welches ein unklares Ziel oder mehrere gegensätzliche Ziele einschließt, die eine Abwägung und Priorisierung erforderlich machen. Nach der oben genannten Definition von komplexen Problemen, auf die sich diese Arbeit bezieht, ist die Phase der Zielausarbeitung jedoch nicht zwingend erforderlich. In der Forschung zum KPL werden den problemlösenden Personen in der Regel konkrete Ziele vorgegeben (Greiff et al., 2016; Wirth & Funke, 2005). Die Phase der (2) Modellbildung und Informationssammlung ist jedoch für KPL elementar. Die in der Definition von KPL angeführte Intransparenz

komplexer Probleme führt dazu, dass die problemlösende Person zwangsläufig Informationen sammeln und in ein mentales Situationsmodell integrieren muss (siehe Kapitel 2.1). Nur so kann das nötige Wissen generiert werden, das zur Lösung eines komplexen Problems erforderlich ist. Der Prozess der Modellbildung und Informationssammlung wird in der Literatur auch häufig als Exploration bezeichnet (Greiff, Molnár, Martin, Zimmermann & Csapó, 2018; Stemmann & Lang, 2018). In der Phase (3) Prognose und Extrapolation wird aus dem mentalen Modell des Problems eine Prognose darüber erstellt, wie sich das Problem in Abhängigkeit der Zeit oder als Konsequenz von Eingriffen in das Problem verändern wird. Dies ist laut Dörner (1989) die Basis dafür, Eingriffe in das Problem nicht nur anhand des gegenwärtigen Zustandes sondern auch anhand der prognostizierten Veränderung auszuwählen. Ist eine Prognose über den zeitlichen Verlauf eines Problems abgeschlossen, so folgt die Phase (4) Planung von Aktionen, Entscheidung und Durchführung. Die Planung und Durchführung von Aktionen sollte eine Problemlösung ganz oder teilweise herbeiführen. Sind mehrere Handlungsalternativen verfügbar, muss außerdem entschieden werden, welche Handlungsalternative umgesetzt werden soll. Wurden die geplanten Handlungen in die Tat umgesetzt, so ist die nächste Phase die (5) Effektkontrolle und Revision der Handlungsstrategien. Ist der erwartete Effekt nicht eingetreten, sollte eine Revision der angewendeten Handlungsstrategien erfolgen. So kann zum Beispiel eine unzureichende Informationssammlung zu einem unvollständigen Modell und somit zu einer falschen Prognose des Systemverhaltens geführt haben. Aber auch die Ausführung geplanter Handlungen, kann fehlerhaft gewesen sein. Je nachdem in welchem Schritt des Prozesses sich ein Fehler identifizieren lässt, kann die problemlösende Person zu diesem Schritt zurückkehren und ihr Vorgehen revidieren.

Das Modell von Dörner (1989) gibt einen guten Überblick über die Prozesse beim KPL. Jedoch kann kritisch angemerkt werden, dass nicht alle Prozesse auf derselben Ebene angesiedelt sind. So wird die Planung und Durchführung von Aktionen konkret im vierten Schritt genannt. Jedoch erfordert auch die Informationssammlung in einem intransparenten und interaktiven System das Planen und Durchführen von Aktionen. Somit muss der Prozess der Informationssammlung als ein Prozess höherer Ordnung betrachtet werden, da er sich aus mehreren untergeordneten Prozessen, wie Planung und Durchführung von Aktionen, zusammensetzt. Dadurch tauchen Planung und Ausführung von Aktionen im Modell implizit doppelt auf, werden jedoch nur an einer Stelle explizit benannt. Auch der sequenzielle Ablauf der einzelnen Prozesse, den Dörners Modell impliziert, kann kritisch betrachtet werden. Zwar betont Dörner, dass von jedem Schritt Rücksprünge zu vorherigen Schritten möglich sind, allerdings ist das Überspringen einzelner Schritte nicht vorgesehen. Es wäre jedoch durchaus denkbar, dass bereits im Anschluss an die Informationssammlung eine Effektkontrolle und Revision der Handlungsstrategien stattfindet, da auch die Informationssammlung strategisch vorgenommen werden sollte (Greiff et al., 2015).

Funke (2003) unterscheidet ähnlich wie Dörner (1989) fünf Phasen, die die einzelnen Prozesse jedoch anders zusammenfassen. Die angenommen Phasen ähneln jedoch denen von Dörner (1989): Die erste Phase ist in beiden Modellen die Zielausarbeitung. Die zweite Phase ist bei Funke (2003) die Hypothesenbildung, die der

Modellbildung, Prognose und Extrapolation in Dörners Modell entspricht. Die Informationssammlung fehlt in Funkes Modell hingegen. Der Hypothesenbildung schließt sich bei Funke die Phase des Planens und Entscheidens an, die sich auch bei Dörner (1989) wiederfindet, wobei Funke die Durchführung der Aktionen nicht explizit in seinem Modell benennt. Nach der Planung und Entscheidung folgen bei Funke (2003) die Überwachung und schließlich die Evaluation als eigenständige Phasen. Bei Dörner (1989) sind Überwachung und Evaluation hingegen zu einer einzigen Phase zusammengefasst. Funke (2003) bezieht diese Phasen jedoch nicht explizit auf KPL, sondern auf Problemlösen im Allgemeinen. Vor diesem Hintergrund lässt sich das Fehlen der Durchführung von Aktionen im Modell erklären.

Allgemein erfüllen handlungstheoretische Ansätze laut Funke (2003) eine ordnende und sinnstiftende Funktion, indem sie die relevanten Prozesse beschreiben. Sie bleiben jedoch auf einer relativ abstrakten Ebene, vor allem wenn es darum geht, Hypothesen über beobachtbare Vorgänge abzuleiten (Greiff, 2012b). Während in Dörners Modell Handlungen zumindest an einer Stelle explizit genannt werden, tauchen diese bei Funke nicht auf. Da die Interaktion, also das aktive Eingreifen der Person, jedoch zentral für die Definition von KPL ist, sollte diese nicht nur in einem Modell der Prozesse beim KPL erwähnt werden, sondern im Zentrum eines solchen Modells stehen. Auch aus empirischer Sicht wäre dies wünschenswert, da sich aus einem solchen Modell Hypothesen über beobachtbares Verhalten ableiten ließen.

Im Gegensatz zu den detaillierten Modellen von Dörner (1989) und Funke (2003) etablierte Funke (2001) eine gröbere Einteilung der Prozesse beim KPL. Diese Einteilung diente in erster Linie dazu, Messinstrumente anhand formaler Kriterien zu erstellen. Im Gegensatz zu den bisher beschriebenen Modellen liegt der Fokus auf der Möglichkeit, beobachtbare Prozesse herzuleiten. Funke (2001) unterschied zu diesem Zweck zwischen den beiden Phasen Wissenserwerb und Wissensanwendung beim KPL, die in vielen Forschungsarbeiten übernommen wurden (Beckmann & Goode, 2017; Blech & Funke, 2005; Greiff et al., 2014). Funke (2001) begründete diese Unterscheidung mit der Intransparenz komplexer Probleme, die es erforderlich mache, zunächst Wissen über das Problem zu generieren, bevor man dieses Wissen zur Lösung des Problems anwenden könne. Der Wissenserwerb wird laut Funke (2001) vollständig abgeschlossen, bevor die Wissensanwendung beginnt. Jedoch ist auch eine Wissensanwendung vor dem Wissenserwerb denkbar, wenn die problemlösende Person über Vorwissen verfügt, das zur Lösung des Problems beiträgt. Darüber hinaus argumentiert Funke (2001), dass Personen auch während der Phase, die eigentlich für den Wissenserwerb vorgesehen ist, versuchen könnten, selbst gesteckte Ziele zu erreichen und dass sie auch während der Wissensanwendung noch Wissen über das Problem hinzugewinnen könnten. Somit scheint eine strenge zeitliche Abfolge von Wissenserwerb und Wissensanwendung zu streng, um generell auf KPL angewendet zu werden.

2.2.2. Ein intentionales Handlungsmodell der Prozesse beim komplexen Problemlösen

Die vorgestellten Modelle haben unterschiedliche Vor- und Nachteile. Die Modelle von Dörner (1989) und Funke (2003) beschreiben detailliert die verschiedenen Prozesse

beim KPL. Allerdings sind sie relativ abstrakt und erlauben kaum die Ableitung von empirisch prüfbaren Hypothesen (Greiff, 2012b). Zudem sind die einzelnen Phasen empirisch nicht präzise zu trennen (Greiff, 2012b). Die gröbere Einteilung von Funke (2001) liefert hingegen empirisch unterscheidbare Phasen, wobei die strenge zeitliche Abfolge infrage gestellt werden kann. Funke (2001) klammert jedoch die Interaktion zwischen Person und Problem vollständig aus, die elementar für die Definition von KPL ist. Um ein theoretisch fundiertes Modell zu erhalten, dass die Interaktion zwischen Person und Problem erklärt sowie Aussagen über beobachtbare Prozesse zulässt, sollen darum die verschiedenen Ansätze zu einem gemeinsamen Modell integriert werden. Das integrierte Modell wird in Abbildung 2-2 dargestellt und im Folgenden näher erläutert.

Funke (2003) stellte die Wichtigkeit der Intention für KPL heraus. KPL kann laut Funke (2003) als das erfolgreiche Umsetzen einer Intention aufgefasst werden. Außerdem sind Handlungen laut Funke (2003) immer mit der Intention einer Person verknüpft. Nach dem Modell von Funke (2001) gibt es jedoch zwei Intentionen, die während des KPL von Bedeutung sind. Im Modell von Dörner gibt es jedoch zwei unterschiedliche Arten von Handlungen: Exploration und zielgerichtetes Verhalten. Als Intentionen dieser beiden Arten von Handlungen können die beiden Phasen von Funke (2001) herangezogen werden: Während Exploration der Intention des Wissenserwerbs dient, dient zielgerichtetes Verhalten der Wissensanwendung. Auch die übrigen Phasen aus den Modellen von Funke und Dörner lassen sich diesen beiden Intentionen zuordnen. Das Ergebnis dieser Zuordnung ist ein intentionales Handlungsmodell, das die Prozesse beim KPL sowohl Intentionen als auch Handlungsphasen zuordnet. Damit greift das Modell die handlungstheoretische Einteilung in Handlungsplanung, Handlungsvollzug und Handlungsbewertung auf, wodurch jede Handlung in einen entsprechenden Kontext gesetzt wird (Greiff, 2012b). Unabhängig davon, welche Intention eine Person gerade verfolgt, steht vor jeder Aktion oder Aktionsfolge ein Planungsprozess. Der Planungsprozess kann die Ausarbeitung von Zielen, eine Prognose oder Extrapolation enthalten (OECD, 2013). Außerdem wird in der Planungsphase die Entscheidung getroffen, welche Aktion(en) als nächstes ausgeführt werden soll(en). Somit erfolgt in diesem Schritt die Festlegung der Intention. Möchte die Person Wissen erwerben, so handelt es sich bei ausgeführten Aktionen um Exploration. Diese Aktionen kommen im Modell von Dörner nur implizit unter dem Punkt „Informationsbeschaffung“ vor. Beim Wissenserwerb muss anschließend an die Exploration die generierte Information verstanden und in ein mentales Modell integriert werden. Außerdem kann die Strategie zum Wissenserwerb reflektiert und ggf. angepasst werden, wenn z.B. nicht die gewünschten Informationen gefunden wurden. Ist die Intention der Person jedoch ihr Wissen anzuwenden, folgen auf die Planung Aktionen, die den aktuellen Zustand dem Zielzustand annähern oder diesen erreichen sollen. Bei der Wissensanwendung folgen auf die Durchführung der Aktionen eine Effektkontrolle sowie gegebenenfalls die Revision der Strategie zur Wissensanwendung. Bei der Effektkontrolle und Revision bei der Wissensanwendung geht es darum zu prüfen, ob das gewünschte Ziel oder Teilziel erreicht wurde und gegebenenfalls das Vorgehen anzupassen. Anders als im Modell von Funke (2001) wird im intentionalen Handlungsmodell nicht davon ausgegangen, dass die Problemlösung mit dem Wissenserwerb begonnen wird. Das Modell lässt auch eine Wissensanwendung vor dem Wissenserwerb zu zum Beispiel in Form von

angewendetem Vorwissen. Zudem kann im intentionalen Handlungsmodell zwischen Wissenserwerb und Wissensanwendung hin- und hergewechselt werden. Beide Intentionen (Wissenserwerb und Wissensanwendung) erfordern also das Planen und Durchführen von Aktionen, eine Kontrolle der Ergebnisse sowie gegebenenfalls die Revision der Strategie.

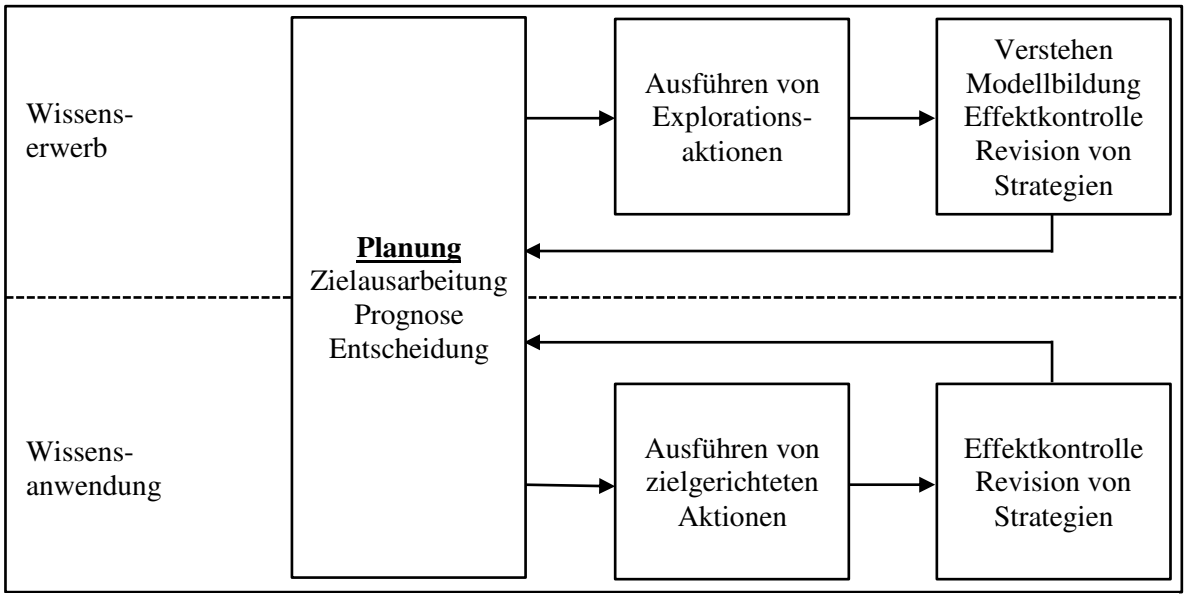

ABBILDUNG 2-2: INTENTIONALES HANDLUNGSMODELL DER PROZESSE BEIM KOMPLEXEN PROBLEMLÖSEN

Das intentionale Handlungsmodell beinhaltet die verschiedenen Phasen aus dem Modell von Dörner (1989) Funke (2003) und Funke (2001). Die Betrachtung von Wissenserwerb und Wissensanwendung als Intentionen gibt den jeweiligen Prozessen einen inhaltlichen Rahmen. Dadurch wird klar benannt, welche beobachtbaren Interaktionen aufgrund welcher Intention auftreten und welche Prozesse diesen vorausgehen und folgen. Das Modell ermöglicht darüber hinaus eine Abgrenzung von analytischem und komplexem Problemlösen auf Prozessebene. Da KPL durch Intransparenz gekennzeichnet ist, ist Wissenserwerb erforderlich, um die benötigten Informationen zur Problemlösung zu erhalten. Beim analytischen Problemlösen sind hingegen alle benötigten Informationen gegeben. Somit muss nur das vorhandene Wissen korrekt angewendet werden. Man könnte also analytisches Problemlösen als den Teil der Wissensanwendung von KPL auffassen. Im Folgenden werden den Implikationen aus dem intentionalen Handlungsmodell empirische Befunde gegenübergestellt.

2.2.3. EMPIRISCHE BEFUNDE

Es gibt verschiedene Untersuchungen, in denen Prozesse beim Lösen von komplexen und nicht-komplexen Problemen beobachtet wurden. Einige Studien deuten darauf hin, dass es beim Problemlösen vorteilhaft ist, Handlungen sorgfältig zu planen (Albert & Steinberg, 2011). Greiff et al. (2016) fanden einen negativen Zusammenhang von der Bearbeitungsfrequenz beim Wissenserwerb und der KPL-Leistung. Dies interpretieren die Autoren als Hinweis auf die Relevanz langsamen (planvollen) Handelns beim Wissenserwerb. Im intentionalen Handlungsmodell ist Planung sowohl für den Wissenserwerb als auch für die Wissensanwendung beim KPL erforderlich. Im Gegensatz dazu steht Planung bei Dörner nur im Zusammenhang mit der Ausführung zielgerichteter

Aktionen (also der Wissensanwendung). Somit macht Dörners Modell keine Aussagen über Planungsprozesse beim Wissenserwerb.

Auch die durchgeführten Interaktionen tauchen nur unvollständig in den Modellen von Dörner (1989)und Funke (2003)auf. Naumann et al. (2014) fanden einen quadratischen Zusammenhang zwischen der Anzahl der durchgeführten Interaktionen und der Leistung beim technologiebasierten, komplexen Problemlösen, wobei die optimale Interaktionsanzahl eine Standardabweichung über dem Durchschnitt lag. Naumann et al. (2014) interpretieren den überwiegend positiven Zusammenhang von Interaktionsanzahl und Leistung dahingehend, dass ein zu passiver Umgang mit computerbasierten KPL-Aufgaben einer hohen Leistung entgegensteht. Dies unterstreicht die Relevanz der Interaktion zwischen Person und Problem beim KPL. Insofern erscheint es wünschenswert, diese Interaktionen auch inhaltlich in Bezug auf ihre Intention einordnen zu können.

Empirische Studien deuten außerdem auf die Wichtigkeit von Exploration beim KPL hin. Je mehr Exploration die Versuchspersonen in einer Studie von Dormann und Frese (1994) zeigten, desto besser war ihre Leistung in einem anschließenden Test. Bell und Kozlowski (2008) konnten zeigen, dass Personen, die während eines Trainings im Umgang mit komplexen Systemen zur Exploration angeregt wurden, eine bessere Leistung in einem anschließenden Test zeigten als Personen, die ein Training ohne diese Anregung erhielten. Diese Ergebnisse deuten im Einklang mit dem intentionalen Handlungsmodell darauf hin, dass Exploration, also die Interaktion zum Wissenserwerb, eine wichtige Rolle beim KPL spielt.

Auch auf die Relevanz einer Reflexion im Anschluss an ausgeführte Handlungen wurde in empirischen Studien bereits hingewiesen. Perez et al. (2017) beobachteten, dass Pausen in der Bearbeitung komplexer Systeme, die nach explorativen Handlungen durchgeführt wurden und somit zur Reflexion geeignet sind, zu einem besseren Verständnis dieser Systeme beitrugen.

Bezogen auf die vorgestellten empirischen Befunde bietet das intentionale Handlungsmodell der Prozesse beim KPL vor Allem den Vorteil, dass es verschiedene Handlungen differenzierbar macht indem diese im Sinne einer Intention interpretierbar werden. Da ausgeführte Aktionen gerade der Teil des Problemlösens sind, der sich direkt beobachten lässt, ist für die empirische Forschung ein Interpretationsrahmen für die ausgeführten Aktionen beim KPL elementar. Das intentionale Handlungsmodell leistet hierzu die explizite Einordnung aller beim KPL auftretenden Handlungen.

2.3. Gruppenunterschiede beim komplexen Problemlösen

Zusätzlich zur Identifizierung relevanter Prozesse beim KPL, ist auch die Untersuchung und Erklärung von Leistungsunterschieden zwischen Gruppen ein Ziel dieser Arbeit. Beim (komplexen) Problemlösen wurden in der Vergangenheit Leistungsunterschiede zwischen verschiedenen Gruppen festgestellt. So wurden in PISA 2012 Leistungsunterschiede beim Problemlösen gefunden (OECD, 2014a). Zum einen wurden Leistungsunterschiede in Abhängigkeit vom Geschlecht festgestellt: Jungen schnitten signifikant besser ab als Mädchen. Ein weiterer Faktor, der zu

Leistungsunterschieden führte, war der Migrationshintergrund: Schülerinnen und Schüler ohne Migrationshintergrund schnitten signifikant besser ab als Schülerinnen und Schüler mit Migrationshintergrund. Auch in Studien, die sich konkret auf KPL beziehen, wurden analoge Leistungsunterschiede zwischen Mädchen und Jungen (Wüstenberg et al., 2014) und zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund (Sonnleitner et al., 2014) festgestellt.

Es gibt bereits Hinweise auf Unterschiede in den KPL-Prozessen, die diese Leistungsunterschiede erklären könnten. Wittmann und Hattrup (2004) beobachteten Unterschiede im Verhalten beim KPL zwischen weiblichen und männlichen Problemlösenden, die Leistungsunterschiede beim KPL erklären könnten. Sie argumentierten, dass männliche Personen eine höhere Leistung beim KPL zeigen als weibliche Personen, da männliche Personen stärker zu risikoreichem Verhalten neigten. Risikoreiches Verhalten bedeutet im Kontext von KPL-Aufgaben, dass stärkere Eingriffe in die Problemsituation vorgenommen werden, wodurch stärkere Veränderungen auftreten, die mehr Informationen über das System preisgeben und dadurch mehr Lerngelegenheiten bieten. Analog dazu stellten Cross, Copping und Campbell (2011) in einer Metaanalyse fest, dass männliche Personen im Allgemeinen ein stärker risikobereites Verhalten als weibliche Personen zeigten. Wüstenberg et al. (2014) beobachteten, dass Jungen häufiger die Vary-one-thing-at-a-time-Strategie (VOTAT) anwendeten. Laut Greiff et al. (2015) ist VOTAT eine Strategie zum Wissenserwerb, bei der einzelne Variablen in einem System verändert werden um ihren Einfluss auf alle anderen beteiligten Variablen zu untersuchen. Das Anwenden dieser Strategie geht oft mit hoher KPL-Leistung einher.

Sonnleitner et al. (2014) untersuchten Verhaltensunterschiede in Abhängigkeit vom Migrationshintergrund. Sie beobachteten, dass Schülerinnen und Schüler mit Migrationshintergrund mehr Exploration zeigten, als Schülerinnen und Schüler ohne Migrationshintergrund. Allerdings zeigten letztere eine höhere KPL-Leistung. Sonnleitner et al. (2014) erklären diese Ergebnisse damit, dass Schülerinnen und Schüler mit Migrationshintergrund Schwierigkeiten mit dem Transfer der entdeckten Informationen in deklaratives Wissen haben. Es scheint also sowohl Leistungsunterschiede als auch Verhaltensunterschiede in Abhängigkeit vom Geschlecht und vom Migrationshintergrund zu geben.

2.4. Messung von komplexem Problemlösen

Die Messung der KPL-Leistung von Personen erfolgt in der Regel computerbasiert. Computerbasierte Tests ermöglichen es im Gegensatz zu papierbasierten Tests, die Dynamik und Intransparenz komplexer Probleme abzubilden (Wirth & Funke, 2005). Die Ansätze um die KPL-Leistungen von Personen zu messen, haben sich im Laufe der Forschung verändert. Im Folgenden werden die am weitesten verbreiteten Ansätze vorgestellt.

2.4.1. Microworlds

Einer der ältesten Ansätze, um KPL-Fähigkeiten computerbasiert zu messen, sind hochkomplexe Simulationen von realen Systemen, die über eine Vielzahl von Variablen gesteuert werden können −auch Microworlds genannt (Blech & Funke, 2005). Eine der

prominentesten Microworlds ist die Simulation „Lohhausen“ (Dörner et al., 1983), in der Versuchspersonen als Bürgermeister*innen einer Kleinstadt agieren sollten. Als solche konnten sie viele verschiedene Variablen wie die Steuerbelastung der Bevölkerung oder den Zinssatz für Einlagen bei der örtlichen Bank beeinflussen, um gemäß der Aufgabenstellung „für das Wohlergehen der Stadt in der näheren und ferneren Zukunft zu sorgen“ (Dörner et al., 1983, S. 107). Lohhausen beinhaltete über 2000 miteinander in Verbindung stehende Variablen, was den Versuchspersonen einen enormen Handlungsspielraum eröffnete. Dörner argumentierte, dass solche hochkomplexen, realitätsnahen Simulationen wie Lohhausen aufgrund ihrer externen Validität besonders geeignet seien, um KPL-Kompetenzen von Personen zu messen.

Allerdings gibt es auch Kritik an diesem Ansatz: Zum einen waren die Ziele der KPL-Aufgaben unklar −es wurde nicht spezifiziert, was genau „das Wohlergehen der Stadt“ (Dörner et al., 1983, S. 107) ausmacht. Somit lag es im Ermessen der Problemlösenden, welche Kriterien sie für eine erfolgreiche Problemlösung heranzogen. Dies erschwerte unter anderem die Vergleichbarkeit der Ergebnisse zwischen Personen. Ein weiterer Kritikpunkt an hochkomplexen Simulationen ist die Messung der Problemlösekompetenz mit nur einer einzigen Aufgabe, was zu Zweifeln an der Reliabilität der Methode führte (Funke, 1988; Greiff & Funke, 2009). Um dieser Kritik zu begegnen, entwickelten verschiedene Forschende Microworlds, die mit weniger Variablen und Handlungsoptionen auskamen (Funke, 1992). Dies führte zu weniger (zeit-)aufwendigen Aufgaben, was den Einsatz mehrerer Aufgaben in einer Testsitzung ermöglichte. Diese Entwicklung gipfelte schließlich im Ansatz der minimalkomplexen Systeme, die häufig in aktueller KPL-Forschung verwendet werden. Der Ansatz der minimalkomplexen Systeme wird in Kapitel 2.4.2 dargestellt.

2.4.2. MINIMALKOMPLEXE SYSTEME

Im Gegensatz zu hochkomplexen Microworlds, die sich durch eine besonders große Anzahl vernetzter Variablen auszeichnen, beinhalten minimalkomplexe Systeme gerade genug Variablen und Beziehungen zwischen diesen, um der Definition von Komplexität zu entsprechen. Minimalkomplexe Systeme können einem von zwei Formalismen zugeordnet werden: 1. Lineare Strukturgleichungssysteme und 2. Finite Automaten (Funke, 2001). Die beiden Aufgabengruppen werden in den folgenden Absätzen näher beschrieben.

Lineare Strukturgleichungssysteme

Bei Aufgaben aus der Gruppe der linearen Strukturgleichungssysteme handelt es sich um Simulationen, die über wenige Eingabevariablen verfügen, mit denen eine geringe Anzahl Ausgabevariablen gesteuert werden kann. Die Wirkbeziehungen zwischen Eingabe- und Ausgabevariablen sind den Problemlösenden zu Beginn der Aufgabe jedoch nicht bekannt. Aufgaben dieser Art sind häufig in eine Wissenserwerbs- und eine Wissensanwendungsphase aufgeteilt (siehe Kapitel 2.2.1). Eine Beispielaufgabe dieses Typs aus dem PISA 2012 Test ist die Climate-Control-Aufgabe. Diese stellt eine Klimaanlage dar, bei der drei Schieberegler verwendet werden können, um die Luftfeuchtigkeit und die Raumtemperatur zu regulieren (Abbildung 2-3). Das erste Teilproblem bestand darin, sich das fehlende Wissen über die Wirkmechanismen des Systems anzueignen. Das zweite Teilproblem war die Erreichung bestimmter Zielwerte für

Luftfeuchtigkeit und Temperatur. Abbildung 2-3 zeigt die Wissenserwerbsaugabe. Diese Aufgabe für sich genommen, erfüllt bereits die Kriterien für ein komplexes Problem: Es gibt einen Anfangszustand (nicht zu wissen, wie die Klimaanlage funktioniert) und ein gewünschtes Ziel (zu wissen, wie die Klimaanlage funktioniert). Zwischen diesen Zuständen liegen Barrieren (es gibt keine Bedienungsanleitung). Um die Barrieren zu umgehen, erfordert das Problem ein aktives Eingreifen (das Manipulieren der Schieberegler). Das System ist insofern intransparent, als dass ohne das aktive Eingreifen das Verhalten des Systems nicht vorhersehbar ist. Durch das Eingreifen verändert sich außerdem der Zustand des Systems dynamisch (Luftfeuchtigkeit und Temperatur steigen oder fallen).

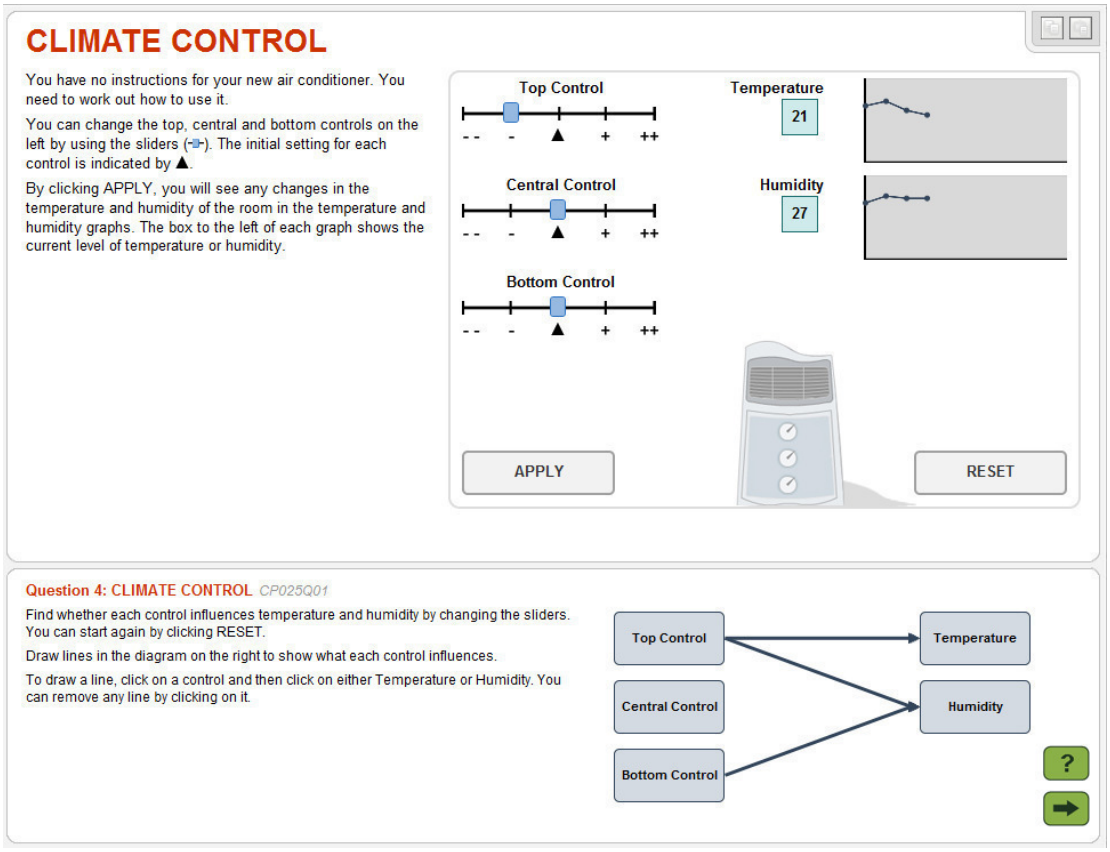

ABBILDUNG 2-3: DIE CLIMATE-CONTROL-AUFGABE AUS PISA 2012 (OECD, 2014A)

Finite Automaten

Finite Automaten lassen sich formal als eine endliche Menge von Zuständen beschreiben, die ein System erreichen kann (Buchner & Funke, 1993; Neubert et al., 2015). Das System kann durch den Einsatz von Operatoren von einem Zustand in einen anderen versetzt werden. Ein Beispiel dieses Aufgabentyps aus dem PISA 2012 Test ist die Tickets-Aufgabe, die einen Fahrkartenautomaten simuliert (Abbildung 2-4). Durch das Betätigen von Buttons kann bei dieser Aufgabe zwischen Fahrkarten für verschiedene Verkehrsmittel und unterschiedlichen Preiskategorien gewechselt werden. Die Kriterien für komplexe Probleme lassen sich auch in der Tickets-Aufgabe feststellen: Es gibt einen Anfangszustand (keine Fahrkarte zu besitzen) und einen gewünschten Zielzustand (eine Fahrkarte zu besitzen). Die Barrieren, die zwischen diesen Zuständen liegen, werden durch die Benutzeroberfläche des Fahrkartenautomaten dargestellt. Das Problem ist insofern

komplex, als dass es intransparent (es ist nicht klar, wie genau verschiedene Tarife gewählt werden und wie teuer diese sind) und dynamisch ist (der Fahrkartenautomat verändert seine Benutzeroberfläche je nach Eingabe). Somit erfordert auch dieses Problem ein aktives Eingreifen, um gelöst zu werden.

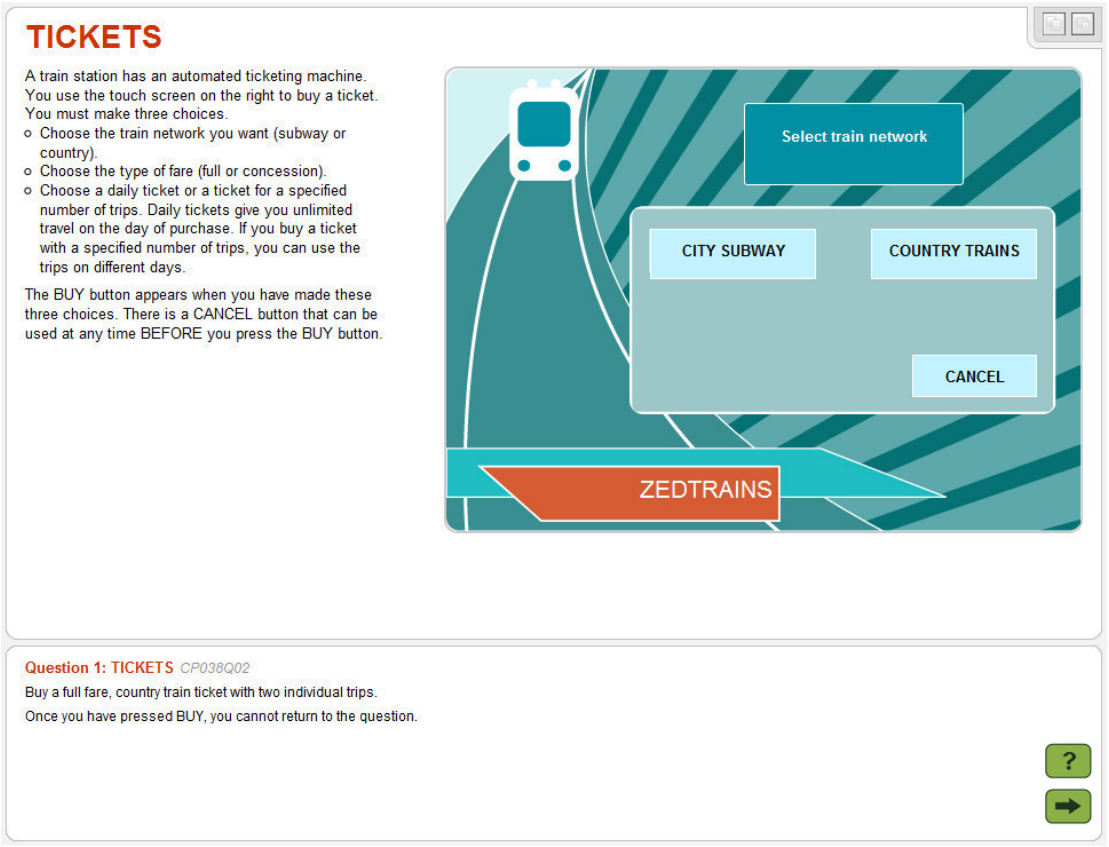

ABBILDUNG 2-4:DIE TICKETS-AUFGABE AUS PISA 2012 (OECD, 2014A)

2.4.3. LOGDATEN

Als Resultat des computerbasierten Messens von KPL-Leistung haben Forschende nicht nur Zugang zu den Ergebnissen der Problemlösetests, sondern auch zu sogenannten Logdaten. Logdaten sind Daten, die bei der Interaktion einer Person mit einem Computer aufgezeichnet werden können. Aus dem alltäglichen Gebrauch von Computern kennt man diese Art der Daten beispielsweise aus der Chronik, die ein Internetbrowser anlegt. In der Chronik zeichnet ein Browser die besuchten Internetseiten und den Zeitpunkt der Besuche auf. Diese Daten werden, von der Person vor dem Bildschirm oft unbemerkt, automatisiert gesammelt und zu einem Verlauf der Internetnutzung an einem bestimmten Gerät zusammengefasst. Bei computerbasierten Assessments können ähnliche Daten erhoben werden. Die Testsysteme können zum Beispiel aufzeichnen, zu welchem Zeitpunkt die getesteten Personen mit welchen Test-Komponenten interagieren. Je nach Testsystem können diese Daten detaillierter (z.B. Aufzeichnen jedes Mausklicks) oder allgemeiner (z.B. Aufzeichnen, wann eine Aufgabe begonnen und beendet wurde) ausfallen. Ursprünglich dienten diese Daten vor allem dazu, Fehler in den Testsystemen nachvollziehen und korrigieren zu können. Da sich mit diesen Daten jedoch auch das Verhalten von Versuchspersonen relativ unkompliziert aufzeichnen lässt, wurde die Nutzung von Logdaten zur Erforschung kognitiver Prozesse in den letzten Jahrzehnten

immer beliebter (Goldhammer et al., 2014; Greiff et al., 2016; Stadler, Fischer & Greiff, 2019).

3. HERLEITUNG DER FORSCHUNGSFRAGEN

Die in Kapitel 2.2 genannten Theorien und empirischen Befunde geben Einblicke in Prozesse, die beim KPL relevant sind. In diesem Kapitel werden aufbauend auf den in Kapitel 2 beschriebenen theoretischen Grundlagen und bisherigen Forschungsergebnissen zu relevanten Prozessen beim KPL Forschungsfragen abgeleitet.

3.1. FORSCHUNGSFRAGE 1: WELCHE PROZESSE HÄNGEN MIT DER LEISTUNG BEIM KOMPLEXEN PROBLEMLÖSEN ZUSAMMEN?

3.1.1. FORSCHUNGSFRAGE 1A: HÄNGT PLANUNG MIT DER LEISTUNG BEIM KPL ZUSAMMEN? (ARBEIT 1)

Sowohl das in Kapitel 2.2.2 vorgestellte intentionale Handlungsmodell als auch ein Reihe weiterer Modelle und Theorien weisen auf die Bedeutsamkeit von Planung beim KPL hin (Dörner, 1989; Leutner et al., 2005; Mayer & Wittrock, 2006). Im intentionalen Handlungsmodell bezieht sich Planung auf die Zielausarbeitung, das Aufstellen von Hypothesen über das Verhalten des Systems (Prognose) und die Auswahl von auszuführenden Aktionen, wobei diese entweder mit der Intention des Wissenserwerb oder der Wissensanwendung erfolgen können. Planung erfüllt also zwei unterschiedliche Funktionen: 1. Planung dient der Zielausarbeitung und der Prognose des Systemverhaltens und ist somit der erste Schritt beim Erstellen einer deskriptiven Repräsentation des Problems. Diese muss jedoch durch Wissenserwerb verifiziert und vervollständigt werden. 2. Planung dient der Auswahl konkreter Handlungen auf Grundlage der vorgenommenen Zielausarbeitung und Prognose. Der Zusammenhang von Planung mit der Leistung beim KPL wurde in Arbeit 1 untersucht.

Da Planungsprozesse nicht direkt beobachtbar sind, werden in der empirischen Forschung Verhaltenskorrelate genutzt, um Planungsprozesse abzubilden (Albert & Steinberg, 2011; Unterrainer & Owen, 2006). Allerdings beziehen sich Studien, die Planungsprozesse untersuchen meist auf analytisches Problemlösen. Albert und Steinberg (2011) untersuchten die Zeit von der Präsentation eines Problems bis zur ersten Aktion der problemlösenden Person als Indikator für Planungsprozesse beim analytischen Problemlösen. Sie fanden heraus, dass Personen umso erfolgreicher waren, je länger sie sich Zeit ließen, bis sie mit der aktiven Bearbeitung begannen. Unterrainer und Owen (2006) stellten außerdem fest, dass sich die Leistung beim analytischen Problemlösen verbesserte, wenn die Versuchspersonen aufgefordert wurden, ihren Lösungsweg zu planen, bevor sie mit der aktiven Bearbeitung begannen. Beide Studien deuten auf den positiven Einfluss initialer Planung beim analytischen Problemlösen hin. Beim KPL könnte eine initiale Planung jedoch weniger nützlich sein. Durch die Intransparenz komplexer Probleme erfordern diese Exploration, um benötigte Informationen zu identifizieren. Ohne diese Informationen kann eine Problemlösung nicht (vollständig) geplant werden. Somit wäre es denkbar, dass beim KPL Planung zu einem späteren oder zu mehreren Zeitpunkten nötig ist. Greiff et al. (2016) beobachteten, dass beim KPL eine niedrigere Interaktionsfrequenz mit höherer Leistung einhergeht. Sie schließen daraus, dass zwischen den einzelnen Interaktionen Planung stattfinden sollte. Allerdings sagen ihre Ergebnisse nichts darüber aus, wie diese Zeit eingeteilt werden sollte, ob wie beim

analytischen Problemlösen initiale Planung vorteilhaft ist oder diese eher später oder zu mehreren Zeitpunkten stattfinden sollte. Ein weiteres Prozessmaß im Zusammenhang mit der Planung ist die aufgewendete Gesamtzeit zur Problemlösung. Goldhammer et al. (2014) fanden einen positiven Zusammenhang zwischen der Gesamtbearbeitungszeit und der Leistung beim KPL. Dieser Zusammenhang war für schwierigere Aufgaben besonders deutlich. Sie argumentieren, dass die Interpretation der Gesamtbearbeitungszeit von der Art der Aufgabe und den benötigten kognitiven Prozessen abhängt. In Aufgaben, die nicht-routiniertes Verhalten erfordern (wie KPL-Aufgaben), kann die Gesamtbearbeitungszeit als ein Zeichen für Anstrengung aufgefasst werden. Es wird davon ausgegangen, dass eine höhere Bearbeitungszeit für höhere Anstrengung spricht. In Aufgaben, die lediglich routiniertes Verhalten erfordern, ist eine hohe Gesamtbearbeitungszeit eher mit Lösungsschwierigkeiten assoziiert. Beim KPL scheint sich die Aufwendung von Zeit für die Aufgabenbearbeitung positiv auszuwirken.

Um den Zusammenhang von Planungsprozessen und Leistung beim KPL zu untersuchen, wurden in Arbeit 1 verschiedene Aspekte von Planung betrachtet. Der erste Aspekt ist die Planungsdauer, die sich in den oben genannten Arbeiten als positiv mit der Leistung beim Lösen (komplexer) Probleme gezeigt hat (Albert & Steinberg, 2011; Greiff et al., 2016; Unterrainer & Owen, 2006). Da initiale Planung jedoch beim KPL nur eingeschränkt möglich ist, wurde stattdessen das längste Zeitintervall zwischen zwei Interaktionen im Problemlöseprozess als Indikator für die Planungsdauer betrachtet. Zusätzlich zur Dauer der Planung stellt sich als Zweites die Frage nach dem Zeitpunkt, zu dem Planung beim KPL erfolgen sollte. Als Indikator für den Zeitpunkt der Planung wurde der Zeitpunkt herangezogen, zu dem das längste Intervall zwischen zwei Interkationen auftritt. Drittens ist die Zeiteinteilung in Planungs- und Handlungsphasen zu betrachten. Dieser Aspekt wurde durch die Varianz aller Intervalle zwischen zwei Interaktionen abgebildet. Mit diesen drei Aspekten von Planung – der Dauer, dem Zeitpunkt und der Zeiteinteilung – wurde der Zusammenhang von Planungsprozessen und Leistung beim KPL untersucht.

3.1.2. Forschungsfrage 1b: Hängt die Anzahl der Interaktionen mit der Leistung beim KPL zusammen? (Arbeit 2, Arbeit 3)

Laut der in Kapitel 2.1 angeführten Definition von KPL ist die Interaktion zwischen Problem und problemlösender Person notwendig, um die für eine erfolgreiche Problemlösung erforderlichen Informationen zu generieren und den gewünschten Zielzustand herbeizuführen. Interaktionen in Form von Exploration und zielgerichteten Aktionen sind auch im Modell in Kapitel 2.2.2 zentral. Die Problemlöseforschung beschäftigt daher unter anderem wie viel Aufwand in Form von Interaktion wünschenswert ist.

Naumann et al. (2014) fanden einen quadratischen Zusammenhang zwischen der Anzahl der durchgeführten Interaktionen und der Leistung beim technologiebasierten, komplexen Problemlösen, wobei die optimale Interaktionsanzahl eine Standardabweichung über dem Durchschnitt lag. Naumann et al. (2014) interpretieren den positiven Zusammenhang von Interaktionsanzahl und Leistung (links vom Optimum) als adäquates Problemlöseverhalten, während der negative Zusammenhang von Interaktionsanzahl und

Leistung (rechts vom Optimum) Desorientierung wiederspiegele. Sie argumentieren, dass sich die meisten Personen beim technologiebasierten Problemlösen zu passiv verhielten, was zum Beispiel eine Folge von Ängstlichkeit beim Umgang mit Technik sein könnte. Da KPL jedoch meist technologiebasiert erfasst wird, könnte auch hier eine zu hohe Passivität einer erfolgreichen Problemlösung im Wege stehen. Andererseits könnte das passive Verhalten der Personen auch eine Folge der Komplexität der Situation sein. Daher wurde in den Arbeiten 2 und 3 untersucht, wie die Anzahl der Interaktionen mit der Leistung beim KPL zusammenhängt.

3.1.3. Forschungsfrage 1c: Hängt Exploration mit der Leistung beim KPL zusammen? (Arbeit 2, Arbeit 3)

Ein weiterer relevanter Prozess, der sich unmittelbar aus dem intentionalen Handlungsmodell ergibt, ist Exploration (vgl. Kapitel 2.2.2). Exploration bezieht sich auf das Identifizieren von Informationen und dient somit dem Wissenserwerb. Da Exploration mit beobachtbaren Interaktionen zwischen Person und Problem einhergeht, ist Exploration theoretisch direkt beobachtbar. Die Schwierigkeit dabei ist jedoch, die eindeutige Klassifizierung von Interaktionen als Exploration. Nach dem intentionalen Handlungsmodell treten Interaktionen zum einen in Form von Exploration beim Wissenserwerb und zum anderen in Form von zielgerichteten Aktionen bei der Wissensanwendung auf. Exploration bezeichnet also die Interaktionen, die nicht zum Ziel haben, den gegenwärtigen in den Zielzustand zu überführen, sondern Informationen über das Problem zu sammeln. Um also zu bewerten, ob eine Interaktion als Exploration gilt oder nicht, sollte ihr Verhältnis zur Problemlösung betrachtet werden. Trägt eine Interaktion nicht zur Problemlösung im Sinne einer Annäherung von aktuellem Zustand und Zielzustand bei, so kann ihr einziger Nutzen darin bestehen, Informationen über das Problem zu generieren. Darüber hinaus können jedoch auch scheinbar zielgerichtete Interaktionen der Informationssammlung dienen, wenn sie im weiteren Verlauf der Problemlösung wiederholt werden. Resultierte das erste Ausführen der zielgerichteten Interaktion nicht in der Lösung des Problems, sondern wurde die gleiche zielgerichtete Interaktion später erneut ausgeführt, so diente das erste Ausführen offenbar der Informationssammlung. Somit ergeben sich zwei unterschiedliche Arten von Exploration: 1. Exploration von lösungsunabhängigen Informationen und 2. Exploration von lösungsrelevanten Informationen (d.h. wiederholt ausgeführte zielgerichtete Aktionen).

Bisherige Forschung hat gezeigt, dass Exploration beim Lernen in komplexen Systemen hilfreich ist (Dormann & Frese, 1994). Laut Bell und Kozlowski (2008) geht Exploration mit einem selbstbestimmten Lernprozess einher. Dieser aktiviere wiederum metakognitive Prozesse wie Planung und Monitoring, die wiederum Lernen und Transfer förderten (Keith & Frese, 2005). Die Ergebnisse dieser Arbeiten zeigen aber vor allem den positiven Effekt von Exploration auf Transferleistungen. Ob Exploration auch für die Lösung des explorierten Problems hilfreich ist, geht aus den Arbeiten nicht hervor. Außerdem wird nicht zwischen lösungsrelevanter und lösungsunabhängiger Exploration differenziert, die sich durchaus unterschiedlich auswirken könnten. Daher wird in den Arbeiten 2 und 3 zum einen der Zusammenhang von Exploration und Erfolg beim KPL

untersucht. Zum anderen befasst sich Arbeit 3 damit, ob die verschiedenen Arten von Exploration unterschiedliche Zusammenhänge mit der Leistung beim KPL aufweisen.

3.1.4. Forschungsfrage 1d: Gibt es komplexe Verhaltensmuster, die mit der Leistung beim KPL zusammenhängen? (Arbeit 3)

Zusätzlich zu den Forschungsfragen 1a-c, die sich mit isolierten Prozessen beschäftigen, wurden in Arbeit 3 zusätzlich komplexe Verhaltensmuster und ihr Zusammenhang mit der Leistung beim KPL analysiert. Dazu wurde ein explorativer Ansatz gewählt. Dieser Ansatz geht über die bisherige Forschung mit Logdaten, die sich vor allem mit Häufigkeiten und Dauern einzelner Prozesse beschäftigte hinaus (siehe Forschungsfragen 1a-c). In der jüngeren Logdaten-Forschung wurden bereits kurze Verhaltenssequenzen in Form von sogenannten n-grams genutzt (He & von Davier, 2015; Stadler, Fischer et al., 2019). Diese bilden jedoch nicht den gesamten Bearbeitungsprozess ab, sondern nur relativ kurze Subsequenzen. Der Vorteil sequenzbasierter Ansätze ist, dass nicht nur einzelne Prozesse analysiert werden können, sondern auch Kombinationen und Sequenzen verschiedener Prozesse. Die Analyse vollständiger Bearbeitungssequenzen erlaubt dabei eine noch elaboriertere Beschreibung der Bearbeitungsprozesse als die Analyse von Subsequenzen. Die Sequenzanalyse basiert auf der Bestimmung der Ähnlichkeiten von Bearbeitungssequenzen verschiedener Personen. Auf Grundlage dieser Ähnlichkeiten können anschließend Cluster von Personen mit ähnlichem Verhalten gebildet werden. Somit kann eine Zuordnung des Vorgehens mit der Leistung beim KPL erfolgen.

Um inhaltlich bedeutsame Verhaltenssequenzen aus den Logdaten zu erzeugen, müssen diese zunächst anhand von geeigneten Kategorien kodiert werden. Die verwendeten Kategorien ergeben sich aus der bisherigen Forschung. Zum einen werden wie in Forschungsfrage 1c zielgerichtete Interaktionen und Exploration unterschieden. Zum anderen erfolgt eine Differenzierung von erstmaligen und wiederholten Interaktionen, sodass vier Verhaltenskategorien entstehen: erstmaliges zielgerichtetes Verhalten, wiederholtes zielgerichtetes Verhalten (lösungsrelevante Exploration), erstmaliges zielunabhängiges Verhalten (erstmalige lösungsunabhängige Exploration) und wiederholtes zielunabhängiges Verhalten (wiederholte lösungsunabhängige Exploration). Zusätzlich zu diesen vier Kategorien wurde das Zurücksetzen der Aufgabe als eigenständige Kategorie betrachtet, da das Zurücksetzen sowohl zielgerichtet als auch zielunabhängig vorgenommen werden konnte. Mithilfe dieser Kodierung wurden komplexe Verhaltensmuster und ihr Zusammenhang mit der Leistung beim KPL in Arbeit 3 untersucht.

3.2. Forschungsfrage 2: Sind Leistungsunterschiede im komplexen Problemlösen zwischen Personen aus verschiedenen sozialen Gruppen durch Prozessunterschiede erklärbar?

Zusätzlich zum Zusammenhang verschiedener Prozesse mit der Leistung beim KPL werden in dieser Arbeit Leistungsunterschiede zwischen soziodemografischen Gruppen untersucht. Auf der Grundlage von bisheriger Forschung wird erwartet,

Leistungsunterschiede beim KPL zwischen Mädchen und Jungen (Wüstenberg et al., 2014) sowie zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund zu finden (Sonnleitner et al., 2014). Dabei steht die Frage im Mittelpunkt, ob sich Leistungsunterschiede zwischen den untersuchten Gruppen durch beim KPL relevante Prozesse erklären lassen.

3.2.1. Forschungsfrage 2a: Gibt es Leistungsunterschiede beim KPL zwischen Mädchen und Jungen, die durch Prozessunterschiede erklärbar sind? (Arbeit 2)

Wittmann und Hattrup (2004) berichteten Verhaltensunterschiede beim KPL in Abhängigkeit vom Geschlecht. Sie beobachteten, dass männliche Versuchsteilnehmer häufiger „riskantes" Verhalten zeigten, was eine stärkere Veränderung der Eingabevariablen meint. Riskante Verhaltensweisen können mit Exploration in Verbindung gebracht werden. Da Exploration laut dem intentionalen Handlungsmodell (siehe Kapitel 2.2.2) auf Prognosen über das Problem basiert, wird dabei stets riskiert Fehler im Sinne von nicht informativen Interaktionen zu begehen. Daher könnte eine vermehrte Exploration seitens der Jungen deren höhere KPL-Leistung erklären. Wüstenberg et al. (2014) verzeichneten bei männlichen Schülern häufiger die Anwendung der VOTAT-Strategie als bei weiblichen. Auch sie beobachteten eine höhere KPL-Leistung bei den Jungen. Die Anwendung der Strategie wurde dann als gegeben betrachtet, wenn jede der Inputvariablen einzeln variiert wurde. Die häufigere Anwendung der Strategie könnte somit jedoch auch aus einer höheren absoluten Interaktionshäufigkeit resultieren. Aus diesem Grund werden sowohl die Explorationshäufigkeit als auch die Anzahl der Interaktionen als mögliche erklärende Faktoren für Leistungsunterschiede zwischen Mädchen und Jungen in Arbeit 2 betrachtet.

3.2.2. Forschungsfrage 2b: Gibt es Leistungsunterschiede beim KPL zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund, die durch Prozessunterschiede erklärbar sind? (Arbeit 2)

Im Gegensatz zum Geschlecht ist die Befundlage zu Verhaltensunterschieden in Abhängigkeit vom Migrationshintergrund weniger eindeutig. Sonnleitner et al. (2014) beobachteten eine geringere KPL-Leistung bei Schülerinnen und Schülern mit Migrationshintergrund. Allerdings beobachteten sie auch, dass Schülerinnen und Schüler mit Migrationshintergrund *mehr* lösungsrelevante Exploration zeigten als Schülerinnen und Schüler ohne Migrationshintergrund – ein Verhalten, dass üblicherweise mit einer hohen KPL-Leistung assoziiert ist. Somit könnte die verstärkte lösungsrelevante Exploration von Schülerinnen und Schülern mit Migrationshintergrund vielleicht sogar noch größere Leistungsunterschiede ausgleichen. In Arbeit 2 wird der Zusammenhang zwischen Migrationshintergrund, Exploration und der Anzahl der Interaktionen genauer untersucht.

4. EMPIRISCHE ARBEITEN

In diesem Kapitel wird zunächst die in den drei Einzelarbeiten verwendete Datengrundlage beschrieben. Anschließend werden die Analysen und Ergebnisse der drei Einzelarbeiten separat dargestellt. Die Diskussion der Ergebnisse erfolgt in Kapitel 5.

4.1. DATENGRUNDLAGE

In den drei Einzelarbeiten wurden jeweils Sekundärdatenanalysen mit Daten aus PISA 2012 durchgeführt. Dazu wurden in den Arbeiten verschiedene Substichproben verwendet, die in Kapitel 4.2 beschrieben werden. Kapitel 4.1.1 stellt hingegen die Gesamtstichprobe dar.

4.1.1. STICHPROBE

Zur Beantwortung der Forschungsfragen wurden Daten des Programme for International Student Assessment (PISA) 2012 herangezogen. Verwendung fanden Logdaten aus dem Problemlösetest, die im Rahmen des Projektes „Prozessindikatoren: Von der Erklärung des Aufgabenerfolgs zum formativen Assessment“ (Profan) von der Organisation for Economic Co-operation and Development (OECD) zur Verfügung gestellt wurden. Außerdem gingen die Fragebogendaten der Schülerinnen und Schüler in Form des Public Use Files der OECD mit ein. Der Einsatz des Problemlösetests erfolgte in 44 Ländern. Daten der Schülerinnen und Schüler aus Zypern konnten jedoch nicht verwendet werden, da die für die Analysen relevanten Angaben aus den Schülerfragebogen im Public Use File nicht vorlagen. Die Gesamtstichprobe für die vorliegende Arbeit bestand somit aus n=82431 Schülerinnen und Schülern aus 43 Ländern (davon 50,05% weiblich und 11,31% mit Migrationshintergrund).

4.1.2. ERHEBUNGSINSTRUMENTE

Problemlösetest

Der KPL-Test in PISA 2012 bestand aus 27 computerbasierten Aufgaben, die in 16 Units organisiert waren und zusammen mit analytischen Problemlöseaufgaben administriert wurden. Jede Unit bestand aus zwei bis drei Aufgaben mit ähnlichem Stimulusmaterial. Die Schülerinnen und Schüler bearbeiteten ein bis zwei von vier verschiedenen Problemlöse-Clustern, die jeweils vier Units umfassten. Die Aufgabenreihenfolge innerhalb der Units war immer gleich. Nach Abschluss einer Aufgabe konnten die Schülerinnen und Schüler nicht mehr zu dieser zurückkehren.

Der computerbasierte KPL-Test erforderte nur grundlegende Computerfähigkeiten wie das Klicken auf virtuelle Schaltflächen und Schieberegler, Drag & Drop und die Eingabe mit der Tastatur. Zu den Antwortformaten gehörten einfache und mehrfache Multiple-Choice-Aufgaben, die durch Anklicken von Radiobuttons beantwortet wurden sowie Aufgaben, bei denen eine Auswahl aus Pulldown-Menüs getroffen werden musste. Des Weiteren gab es Aufgaben, bei denen Teile von Diagrammen gezeichnet werden mussten und Aufgaben mit offenen Textantworten (OECD, 2013). Vor dem Test wurde ein Tutorial durchgeführt, mit dem die Schülerinnen und Schüler die erforderlichen Fähigkeiten üben konnten. Für eine detaillierte Erläuterung der Aufgaben siehe OECD (2013).

Schülerfragebogen

Mit dem Schülerfragebogen wurden Angaben bezüglich der Voraussetzungen und Prozesse des Lernens (z.B. demografische Angaben, Angaben zur Schule und zum Elternhaus) sowie nicht-kognitive Lernergebnisse wie Einstellungen, Motive, Interessen und Überzeugungen abgefragt (OECD, 2014b).

4.1.3. Prozedur

In PISA 2012 wurden die Kompetenzdomänen Mathematik, Lesen, Naturwissenschaften, Problemlösen und Finanzkompetenz erhoben. Der Problemlösetest war neben Mathematik und Lesen Teil des optionalen computerbasierten Tests, an dem 44 Länder freiwillig teilnahmen. Den Schülerinnen und Schülern wurden zwei computerbasierte Aufgaben-Cluster vorgelegt, von denen keines, eines oder beide Problemlösecluster waren. In die Analysen dieser Arbeit sind jedoch nur Daten von Schülerinnen und Schülern eingegangen, die mindesten ein Problemlöse-Cluster bearbeiteten. Die computergestützte Testung erfolgte nach den papierbasierten PISA-Tests. Die Schüler hatten pro Cluster 20 Minuten Zeit für die Aufgabenbearbeitung (OECD, 2014c).

4.1.4. Datenaufbereitung

Das PISA-Scoring der KPL-Aufgaben wurde in der vorliegenden Arbeit nicht verwendet, da es teilweise auf dem Verhalten der Schülerinnen und Schüler in der jeweiligen Aufgabe basierte. So wurde beispielsweise eine richtige Antwort bei einigen Aufgaben nur dann als vollständig richtig bewertet, wenn zuvor alle relevanten Antwortalternativen erkundet wurden. Dadurch sollte sichergestellt werden, dass richtig geratene Lösungen einen niedrigeren Score bekamen als vollständig erarbeitete Lösungen. Da in dieser Arbeit jedoch der Zusammenhang zwischen Prozessen, die während der Aufgabe ablaufen und der Leistung in der Aufgabe untersucht werden soll, wurde ein dichotomes Scoring verwendet, das ausschließlich angibt, ob eine Aufgabe richtig oder falsch gelöst wurde. Aufgaben, zu denen keine Lösung abgegeben wurde, erhielten ebenfalls die Kodierung falsch. Dieses dichotome Scoring wurde auf Grundlage der Logdaten vorgenommen.

4.2. Analysen und Ergebnisse

Im Folgenden werden die in den Einzelarbeiten durchgeführten Analysen sowie deren Ergebnisse zusammengefasst. Eine ausführliche Darstellung der Analysen und Ergebnisse findet sich in den Einzelarbeiten. Die Ergebnisse der Einzelarbeiten werden im Kapitel 5 diskutiert.

4.2.1. Arbeit 1: Die Rolle von Planung beim komplexen Problemlösen

In der ersten Arbeit wurde der Zusammenhang von Planungsaktivitäten im Prozess der komplexen Problemlösung mit dem Aufgabenerfolg untersucht (Forschungsfrage 1a). Dazu erfolgte die Bildung von drei Prozessindikatoren aus den Logdaten, die unterschiedliche Aspekte von Planungsverhalten abbilden sollen. Aufbauend auf bisherigen Erkenntnissen wurde untersucht, wie sich der Zeitpunkt und die Dauer von Planungsphasen sowie die Variation der Dauer von Planungsphasen im

Problemlöseprozess auf den Aufgabenerfolg auswirken. Zudem wurde überprüft, ob die Ausprägung der Effekte aufgabenabhängig ist. Inwieweit ein Interaktionseffekt der drei Prozessindikatoren auf den Aufgabenerfolg besteht, war ebenfalls Bestandteil der Betrachtung.

Als Indikator für die Dauer von Planung wurde die Dauer des längsten aufgetretenen Intervalls zwischen zwei aufeinanderfolgenden Interaktionen genutzt. Als Indikator für den Zeitpunkt von Planung diente der Zeitpunkt des Längsten aufgetretenen Intervalls zwischen zwei aufeinanderfolgenden Interaktionen. Als Indikator für die Variation der Dauer von Planungsintervallen wurde die Varianz aller Intervalldauern zwischen jeweils zwei aufeinanderfolgenden Interaktionen betrachtet. Der Zusammenhang der drei Indikatoren mit dem Aufgabenerfolg, die Interaktionseffekte zwischen den Indikatoren und die aufgabenspezifischen Effekte wurden mit Hilfe von generalisierten linearen Mischmodellen geschätzt. Modellvergleiche mithilfe des Likelihood-Ratio-Tests wurden verwendet, um die aufgabenspezifischen Effekte auf Signifikanz zu überprüfen. Die Analysen erfolgten mit der deutschen Substichprobe aus PISA 2012 (n=1346).

Die Ergebnisse zeigen, dass kein allgemeiner Effekt der Dauer des längsten Planungsintervalls (β=-0,05, p=0,639) oder der Variation der Dauern der Planungsintervalle (β=-1,85, p=0,155) auf die Lösungswahrscheinlichkeit beim komplexen Problemlösen bestand. Der Zeitpunkt der Planung hatte jedoch einen negativen Effekt auf die Lösungswahrscheinlichkeit der Aufgabe (β=-0,13, p=0,005). Außerdem waren aufgabenspezifische Effekte für alle drei Indikatoren zu beobachten. Je schwieriger die Aufgabe, desto negativer wirkt sich die Dauer des längsten Planungsintervalls auf die Lösungswahrscheinlichkeit aus (r=-0,04). Der Likelihood-Ratio-Test zeigte eine signifikant bessere Passung des Modells mit dem aufgabenspezifischen Effekt als ohne den aufgabenspezifischen Effekt (χ^2=46,61, p<0,001). Je leichter die Aufgabe, desto negativer wirkt sich eine späte Planung aus (r=-0,51). Bei schwierigen Aufgaben ist der Effekt hingegen eher klein. Auch in diesem Fall deutet der Likelihood-Ratio-Test auf einen signifikanten Effekt hin (χ^2=16,20, p<0,001). Der insgesamt negative Effekt der Variation kehrt sich für schwere Aufgaben um und wird positiv (r=-0,39). Auch hier deutet der Likelihood-Ratio-Test auf einen signifikanten aufgabenspezifischen Effekt hin (χ^2=34,26, p<0,001). Zudem liegen signifikante Interaktionseffekte zwischen Dauer und Zeitpunkt (β=0,18, p=0,013), zwischen Zeitpunkt und Variation (β=-5,56, p=0,020) und zwischen Dauer, Zeitpunkt und Variation vor (β=2,73, p=0,036).

Die Effekte der drei untersuchten Indikatoren scheinen somit allesamt von der Aufgabe abzuhängen. Bei leichteren Aufgaben konnten – ähnlich wie bei analytischen Problemen – positive Effekte langer, initialer Planungsintervalle beobachtet werden (Unterrainer & Owen, 2006). Mit steigender Aufgabenschwierigkeit veränderten sich jedoch die Effekte der Indikatoren. Bei schwierigen Aufgaben bestand kein positiver Effekt einer Planung zu Beginn des Bearbeitungsprozesses. Eine mögliche Erklärung dafür ist, dass schwierige Aufgaben Exploration erfordern, um Informationen zu generieren, bevor eine umfassende Planung möglich ist. Außerdem erforderten schwierige Aufgaben im Gegensatz zu leichten Aufgaben eine Einteilung der Zeit in Phasen höherer und niedrigerer Aktivität. Dies könnte darauf hindeuten, dass Planung nicht wie bei leichten

Aufgaben zu Beginn, sondern verteilt über mehrere Intervalle erfolgen sollte zwischen denen Phasen höherer Aktivität (z.B. für Exploration) stattfinden. Außerdem besteht eine hohe gegenseitige Abhängigkeit der Effekte. So wirkte sich die ungleichmäßige Zeiteinteilung nur positiv aus, wenn der Großteil der Planung zu einem frühen Zeitpunkt und in einem eher kurzen Intervall erfolgt. Auch der positive Effekt der frühen Planung, der bei leichten Aufgaben beobachtet wurde, bestand nur bei relativ kurzen, ungleichmäßigen Planungsintervallen.

4.2.2. Arbeit 2: Die Erklärung von Gruppenunterschieden beim komplexen Problemlösen anhand von Prozessdaten

Ziel der zweiten Arbeit war es, Interaktions- und Explorationshäufigkeit beim KPL zu untersuchen (Forschungsfrage 1b und 1c). Hinzu kam die Erklärung von Leistungsunterschieden zwischen gesellschaftlichen Gruppen durch Prozessunterschiede (Forschungsfrage 2). Untersucht wurden Leistungsunterschiede zwischen Mädchen und Jungen (Forschungsfrage 2a) sowie zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund (Forschungsfrage 2b). Zu diesem Zweck wurden zwei Indikatoren für Prozessindikatoren gebildet: Zum einen wurde die Anzahl der Interaktionen verwendet, die bereits von Naumann et al. (2014) beim technologiebasierten Problemlösen untersucht wurde. Zum anderen sollte zwischen zielgerichtetem und explorativem Verhalten unterschieden werden, indem die Anzahl der Explorationsschritte ebenfalls als Indikator verwendet wurde. Als Explorationsschritte wurden alle Interaktionen verstanden, die nicht unmittelbar zur Aufgabenlösung führten. Es wurde also nicht zwischen lösungsrelevanter und lösungsunabhängiger Exploration unterschieden.

Für die Analysen wurden die Daten aller Schülerinnen und Schüler aus PISA 2012 verwendet, die mindestens ein KPL-Aufgabencluster bearbeiteten mit Ausnahme der Schülerinnen und Schüler aus Zypern, da von diesen die Angaben zu Geschlecht und Migrationshintergrund in den Fragebogendaten nicht vorlagen. Außerdem wurden die Daten aus Korea nicht verwendet, da für den koreanischen Datensatz nicht alle geschätzten Modelle konvergierten. Somit enthielt die Stichprobe n=81.039 Schülerinnen und Schüler aus 42 Ländern. Pro Land wurden vier Mediationsmodelle geschätzt, wobei der Prädiktor (Geschlecht/Migrationshintergrund) sowie die Mediatorvariable (Anzahl der Interaktionen/Anzahl der Explorationsschritte) variiert wurden. Lesefähigkeit wurde dabei in Form von Weighted Likelihood Estimators auf der Basis des Print-Lesetests als Kontrollvariable verwendet, um auszuschließen, dass unterschiedliche sprachliche Fähigkeiten der betrachteten Gruppen die Ergebnisse verzerren. Die Ergebnisse der entsprechenden Modelle wurden anschließend metaanalytisch über die Länder hinweg zusammengefasst, wobei jedes Land als „Studie“ in die Metaanalyse einging.

In den Ergebnissen zeigten sich die erwarteten Leistungsunterschiede zugunsten der Jungen (β_{total}=-0,28, p<0,001) und zugunsten der Schülerinnen und Schüler ohne Migrationshintergrund (β_{total}=-0,16, p=0,001 bzw. β_{total}=-0,12, p=0,003). Die beiden Prozessindikatoren Anzahl der Interaktionen (β=0,71, p<0,001 bzw. β=0,74, p<0,001) und Anzahl der Explorationsschritte (β=0,44, p<0,001 bzw. β=0,44, p<0,001) wiesen beide einen positiven Zusammenhang mit der KPL-Leistung auf. Die Anzahl der Interaktionen ($\beta_{indirekt}$=-0,19, p<0,001, κ^2=0,13) konnte den Geschlechtereffekt teilweise erklären. Die

Anzahl der Explorationsschritte ($\beta_{indirekt}$=-0,23, p<0,001, κ^2=0,17) erklärte den Geschlechtereffekt hingegen vollständig. Der Effekt des Geschlechts auf die KPL-Leistung reduzierte sich unter Kontrolle der Anzahl der Interaktionen; er blieb aber weiterhin signifikant (β_{direkt}=-0,08, p<0,001). Unter Kontrolle der Anzahl der Explorationsschritte reduzierte sich der direkte Effekt des Geschlechts auf die Leistung noch stärker und war nicht mehr signifikant (β_{direkt}=0,03, p=0,382). Der Leistungsunterschied zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund konnte weder durch die Anzahl der Interaktionen ($\beta_{indirekt}$=-0,07, p=0,077) noch durch die Anzahl der Explorationsschritte ($\beta_{indirekt}$=0,02, p=0,235) erklärt werden. Der Effekt des Migrationshintergrundes auf die KPL-Leistung war unter Berücksichtigung der Anzahl der Interaktionen jedoch nicht mehr substanziell (β_{direkt}=-0,07, p=0,146). Unter Berücksichtigung der Anzahl der Explorationsschritte blieb der Effekt substanziell (β_{direkt}=-0,15, p<0,001).

Die Ergebnisse zeigen, dass sowohl die Anzahl aller Interaktionen als auch die Explorationshäufigkeit positiv prädiktiv für die Leistung beim KPL sind. Der Leistungsunterschied zwischen Mädchen und Jungen wird durch die Explorationshäufigkeit jedoch stärker mediiert. Somit kann der Geschlechterunterschied tatsächlich durch direkt beobachtbare Verhaltensunterschiede abgebildet werden: Unter Kontrolle der Explorationshäufigkeit gibt es keinen Leistungsunterschied beim KPL zwischen Mädchen und Jungen. Der Leistungsunterschied zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund lässt sich hingegen nicht mit den hier untersuchten Prozessindikatoren erklären.

4.2.3. Arbeit 3: Exploration von Verhaltensmustern beim komplexen Problemlösen

Das Ziel der dritten Arbeit war es, die Rolle von Explorationsverhalten beim KPL näher zu beleuchten (Forschungsfrage 1c) und komplexe Verhaltensmuster zu identifizieren, die mit Erfolg oder Misserfolg beim KPL assoziiert sind (Forschungsfrage 1d). Im Unterschied zu den Analysen aus Arbeit 1 und 2 wurde ein explorativer Ansatz verwendet, der sich nicht nur auf die Häufigkeiten oder Dauern einzelner Verhaltensweisen bezieht, sondern die Abfolge verschiedener Verhaltensweisen berücksichtigt (Gabadinho, Ritschard, Müller & Studer, 2011). Dazu wurden die einzelnen Logevents in den Daten zunächst anhand inhaltlich bedeutsamer Kategorien kodiert. Die Kodierung der Logevents erfolgte ähnlich wie in Arbeit 2 entweder als zielgerichtet (alle Interaktionen, die Teil der kürzesten möglichen Problemlösung sind) oder Exploration (alle Interaktionen, die *nicht* Teil der kürzesten korrekten Problemlösung sind). Einzig das Zurücksetzen einer Aufgabe konnte keiner der beiden Kategorien eindeutig zugeordnet werden, da es sowohl Teil einer effizienten Lösung als auch Teil der lösungsunabhängigen Exploration sein kann. Im Unterschied zu der Kodierung in Arbeit 2 wurden wiederholte zielgerichtete Interaktionen (lösungsrelevante Exploration) und lösungsunabhängige Exploration unterschieden. Exploration, die zur Zielerreichung erforderlich war, wurde somit von lösungsunabhängiger Exploration unterschieden. Die Logevents wurden demnach anhand der folgenden fünf Kategorien kodiert: erstmaliges zielgerichtetes Verhalten, wiederholtes zielgerichtetes Verhalten (lösungsrelevante Exploration), erstmalige lösungsunabhängige

Exploration, wiederholte lösungsunabhängige Exploration und Zurücksetzen. Somit ergab sich für jede Person eine Sequenz, die aus diesen fünf Kategorien bestand.

Die Analysen wurden mit zwei Aufgaben durchgeführt. Eine Aufgabe aus dem Framework der linearen Strukturgleichungssysteme (Climate-Control-Aufgabe, siehe Abbildung 2-3) und eine aus dem Framework der finiten Automaten (Tickets-Aufgabe, siehe Abbildung 2-4). Für die Analysen in Arbeit 3 wurden die Logdaten von allen Schülerinnen und Schülern verwendet, die die beiden untersuchten KPL-Aufgaben bearbeiteten. Dies führte zu einer Stichprobe von 30089 Schülerinnen und Schülern aus 42 Ländern. Die Analysen wurden jeweils für beide Aufgaben und für richtige und falsche Antworten getrennt durchgeführt. Nach der Kodierung der Logdaten wurde die Ähnlichkeit der Verhaltenssequenzen der Schülerinnen und Schüler mittels Optimal Matching bestimmt (Abbott & Forrest, 1986). Anschließend wurde eine Clusteranalyse auf Grundlage der Ähnlichkeit durchgeführt, die zu Clustern ähnlichen Verhaltens führte (see Studer, 2013). Um die optimale Anzahl von Clustern zu ermitteln, wurden verschiedene Gütekriterien herangezogen (normalisierte punktbiseriale Korrelation, Average Silhouette Width und Huberts C-Index, s. Studer, 2013). Zusätzlich wurde die allgemeine KPL-Leistung der Schülerinnen und Schüler bestimmt, indem Weighted Likelihood Estimators (WLEs) auf Grundlage der Antworten in allen KPL-Aufgaben außer den beiden untersuchten gebildet wurden (s. Robitzsch, Kiefer & Wu, 2019). Somit konnte das Verhalten der Schülerinnen und Schüler nicht nur mit der unmittelbaren Lösung des betreffenden Problems, sondern auch mit einem Maß für die allgemeine KPL-Leistung in Beziehung gesetzt werden.

Es wurden je nach Aufgabe und je nach Richtigkeit der Lösung unterschiedlich viele Cluster gefunden. Einige Verhaltensmuster zeigten sich jedoch in beiden Aufgaben. In beiden Aufgaben wiesen die nicht erfolgreichen Verhaltenssequenzen deutlich mehr lösungsunabhängige Exploration auf als die erfolgreichen Verhaltenssequenzen (Climate-Control-Aufgabe: χ^2(erstmalige zielunabh. Exploration)=5,57, p=0,018, χ^2(wiederholte zielunabh. Exploration)=5,75, p=0,017; Tickets-Aufgabe: χ^2(erstmalige zielunabh. Exploration)=10,26, p=0,001, χ^2(wiederholte zielunabh. Exploration)=7,32, p=0,007). In Bezug auf zielgerichtetes Verhalten gab es nur für wiederholtes zielgerichtetes Verhalten, also für lösungsrelevante Exploration, in der Tickets-Aufgabe einen signifikanten Unterschied (Climate-Control-Aufgabe: χ^2(erstmalige zielger. Verhalten)=3,50, p=0,061, χ^2(wiederholte zielger. Verhalten)=3,53, p=0,060; Tickets-Aufgabe: χ^2(erstmalige zielger. Verhalten)=0,05, p=0,821, χ^2(wiederholte zielger. Verhalten)=6,82, p=0,009). Die höchste allgemeine Problemlöseleistung zeigten in beiden Aufgaben Schülerinnen und Schüler, die die längsten Verhaltenssequenzen mit überwiegend zielgerichtetem Verhalten zeigten. Diese Schülerinnen und Schüler gehen über das notwendige Maß an Interaktionen hinaus und scheinen ihre Antwort mehrfach zu verifizieren (d.h. viel lösungsrelevante Exploration durchzuführen), bevor sie sich auf eine Antwort festlegen. Eine etwas geringere allgemeine KPL-Leistung zeigten in der Tickets-Aufgabe Schülerinnen und Schüler, die ein eher minimalistisches Vorgehen an den Tag legten (also die kürzeste oder nahezu kürzeste Verhaltenssequenz, die zum Erfolg führt). In der Climate-Control-Aufgabe gab es keinen signifikanten Leistungsunterschied zwischen minimalistischem Vorgehen und einem

gemischten Ansatz aus zielgerichtetem Verhalten und lösungsunabhängiger Exploration. In der Tickets-Aufgabe gab es außerdem Cluster von Schülerinnen und Schülern, die die richtige Antwort errieten. Gingen die Schülerinnen und Schüler dabei zielgerichtet vor, so zeigte sich eine höhere allgemeine KPL-Leistung, als wenn die geratene Lösung offenbar zufällig zustande kam.

Zielgerichtetes Raten trat in der Tickets-Aufgabe auch in den falschen Lösungen auf. Die allgemeine Problemlöseleistung dieser Gruppe war in der Tickets-Aufgabe vergleichbar mit Schülerinnen und Schülern, die überwiegend zielgerichtetes Verhalten zeigten, die Aufgabe aber dennoch falsch lösten. In der Climate-Control-Aufgabe war es ebenfalls die Gruppe der Schülerinnen und Schüler mit dem überwiegend zielgerichteten Verhalten, die die höchste allgemeine KPL-Leistung zeigte (verglichen mit den anderen Gruppen, die die Aufgabe ebenfalls falsch lösten). In beiden Aufgaben gab es Gruppen, die relativ lange von lösungsunabhängiger Exploration dominierte Sequenzen zeigten. Diese Gruppen wiesen eine noch geringere allgemeine KPL-Leistung auf. Die niedrigste allgemeine KPL-Leistung zeigte sich jedoch in Gruppen mit extrem kurzen Verhaltenssequenzen und hohem lösungsunabhängigen Explorationsanteil. Schülerinnen und Schüler in diesen Gruppen haben die Lösung vermutlich geraten, wobei sie im Gegensatz zu den vorher genannten Gruppen kaum oder gar nicht zielgerichtet vorgegangen sind, oder sie haben die Aufgabe nicht beantwortet.

Häufigste Schwierigkeiten beim KPL sind also das frühzeitige Beenden der Bearbeitung und die Ablenkung durch lösungsunabhängige Inhalte. Die Ergebnisse waren ähnlich in den beiden untersuchten KPL-Aufgaben, jedoch traten die jeweiligen Verhaltensmuster in den beiden Aufgaben unterschiedlich häufig auf. Dies deutet darauf hin, dass ähnliche Verhaltensweisen in KPL-Aufgaben aus den beiden untersuchten Frameworks mit Erfolg bzw. Misserfolg assoziiert sind.

5. DISKUSSION

In der vorliegenden Arbeit wurden verschiedene Prozesse des komplexen Problemlösens sowie ihr Zusammenhang mit der Leistung beim KPL untersucht. Es wurde ein theoretisches Modell aus vorherigen Modellen abgeleitet, das die kontextuelle Einordnung der Prozesse ermöglicht. Weiterhin konnte gezeigt werden, dass die zentralen Prozesse der Planung, Exploration, sowie zielgerichtetes Verhalten mit der Leistung beim KPL zusammenhängen. Außerdem wurden Leistungsunterschiede im KPL in Abhängigkeit vom Geschlecht und vom Migrationshintergrund untersucht. In den folgenden Abschnitten werden die Ergebnisse der drei Einzelarbeiten zusammenfassend diskutiert.

5.1. ZUSAMMENHANG VON PROZESSEN UND ERFOLG BEIM KPL (FORSCHUNGSFRAGE 1)

Zur Beantwortung von Forschungsfrage 1 wurden die Zusammenhänge verschiedener kognitiver Prozesse mit der Leistung beim KPL untersucht. Arbeit 1 beschäftigte sich mit Planungsprozessen, während sich die Arbeiten 2 und 3 mit Interaktions- und Explorationshäufigkeiten befassten. In Arbeit 3 wurden außerdem explorativ komplexe Verhaltensmuster beim KPL untersucht. Die Ergebnisse werden nachfolgend im Kontext der entsprechenden Forschungsfragen diskutiert.

5.1.1. DIE ROLLE VON PLANUNG (FORSCHUNGSFRAGE 1A)

Der Frage nach der Rolle von Planungsprozessen beim KPL wurde in Arbeit 1 nachgegangen. Dazu wurden drei verschiedene Aspekte von Planungsverhalten betrachtet, die unterschiedlich mit der KPL-Leistung zusammenhingen. Alle drei Aspekte von Planung zeigten aufgabenspezifische Effekte. Nur ein Aspekt, der Zeitpunkt der Planung, zeigte zusätzlich einen aufgabenunabhängigen Effekt. Es wurden außerdem Interaktionseffekte der verschiedenen Planungsaspekte auf die KPL-Leistung gefunden.

Die aufgewendete Planungsdauer zeigte einen aufgabenabhängigen Effekt auf die Leistung beim KPL. Während bei einigen Aufgaben eine lange Planungsdauer hilfreich war, erwies sich dies bei anderen Aufgaben als eher hinderlich. Die Aufgabenschwierigkeit schien jedoch nicht das entscheidende Aufgabenmerkmal zu sein, dass zu diesem Unterschied führt. Ein möglicherweise ursächliches Aufgabencharakteristikum könnte hingegen die Komplexität der Aufgabe sein, die Beckmann und Goode (2017) als Anzahl gleichzeitig zu verarbeitender Informationen verstehen (OECD, 2013). Aufgabeneigenschaften wie diese könnten den erforderlichen Planungsaufwand pro Interaktion oder Interaktionsfolge und somit die Förderlichkeit der Planungsdauer beeinflussen. Diese Interpretation ist auch im Einklang mit Ergebnissen von Goldhammer, Naumann und Greiff (2015). Diese fanden einen aufgabenabhängigen Effekt der Bearbeitungszeit bei Aufgaben zum schlussfolgernden Denken. Hier gab es in schwierigen Aufgaben einen positiven Effekt der aufgewendeten Zeit, während der Effekt bei leichteren Aufgaben negativ war. Bei diesen Aufgaben ist die gleichzeitig zu verarbeitende Menge an Informationen schwierigkeitsbestimmend (vgl. Neubauer, 1990).

Der Effekt des Zeitpunktes der Planung auf die KPL-Leistung zeigt, dass Planung idealerweise möglichst früh im Bearbeitungsprozess stattfinden sollte. Jedoch zeigte sich auch hier ein aufgabenabhängiger Effekt. Bei leichten Aufgaben ist eine frühe Planung vorteilhaft. Bei schwereren Aufgaben hat eine frühe Planung hingegen keinen Vorteil. Ein möglicher Grund für diesen Zusammenhang ist, dass bei leichten Aufgaben zu einem früheren Zeitpunkt vorausgeplant werden kann. Bei schwierigen Aufgaben ist dies möglicherweise aufgrund der höheren Intransparenz und daraus resultierenden Erforderlichkeit von Exploration nicht möglich. Somit scheint sich der Zeitpunkt der Planung bei leichten KPL-Aufgaben ähnlich wie bei analytischen Problemen auszuwirken. Albert und Steinberg (2011) und Unterrainer und Owen (2006) fanden, dass sich in analytischen Problemlöseaufgaben eine lange Planung zu Beginn des Prozesses positiv auswirkte. Schwierigere KPL-Aufgaben erfordern hingegen möglicherweise zunächst eine umfassende Exploration, bevor Planung sinnvoll stattfinden kann. Diese Interpretation wird auch von dem in Kapitel 2.2.2 vorgestellten intentionalen Handlungsmodell gestützt. Demnach wird im Verlauf der Aufgabenbearbeitung die Planung aufwändiger, da das mentale Modell im Verlauf der Aufgabenbearbeitung erweitert wird. Dadurch wird wiederum die Prognose des Systemverhaltens aufwendiger und erfordert mehr Zeit, da mehr Informationen berücksichtigt werden müssen. Bei analytischen Problemen muss das mentale Modell hingegen nicht erst durch aktiven Wissenserwerb gebildet werden. Daher kann bei analytischen Problemen die Planung gebündelt zu Beginn stattfinden.

Auch die Zeiteinteilung hatte einen aufgabenabhängigen Effekt, der mit der Aufgabenschwierigkeit zusammenhing. In leichten Aufgaben scheint es von Vorteil zu sein, die Zeit gleichmäßig einzuteilen. In schwierigen Aufgaben sind hingegen Phasen höherer und niedrigerer Aktivität vorteilhaft. Dieser Zusammenhang mit der Aufgabenschwierigkeit könnte daraus resultieren, dass bei schwierigen Aufgaben möglicherweise komplexere Aktionsfolgen erforderlich sind, die jeweils mehr Planung erfordern als einzeln durchgeführte Schritte, die für die Lösung leichter Aufgaben genügen.

5.1.2. Die Rolle der Interaktionshäufigkeit (Forschungsfrage 1b)

Der Frage nach der Rolle der Interaktionshäufigkeit beim KPL wurde in den Arbeiten 2 und 3 nachgegangen. Während in Arbeit 2 der Zusammenhang von Interaktionshäufigkeit und Problemlöseleistung direkt untersucht wurde, kam in Arbeit 3 ein explorativer Ansatz zur Anwendung, um komplexe Verhaltensmuster beim KPL zu analysieren, die sich auch hinsichtlich der Interaktionshäufigkeit unterschieden.

In Arbeit 2 wurde anhand der kompletten KPL-Stichprobe von PISA 2012 der Zusammenhang von Interaktionshäufigkeit und KPL-Leistung untersucht. Es zeigte sich ein positiver Zusammenhang der Interaktionshäufigkeit mit der KPL-Leistung in sämtlichen teilnehmenden Ländern. Die Effekte variierten lediglich in ihrer Stärke, jedoch nicht in Richtung. Dieses Ergebnis stimmt mit den Ergebnissen von Naumann et al. (2014) überein, die ähnliche Effekte beim computerbasierten Problemlösen feststellten. In Arbeit 3 wurden bei zwei ausgewählten KPL-Aufgaben vollständige Verhaltenssequenzen untersucht. Dabei fiel auf, dass auch hier die längeren Sequenzen bei richtigen Antworten

beobachtet wurden. Außerdem wiesen Schülerinnen und Schüler, die längere Sequenzen zeigten, höhere allgemeine KPL-Fähigkeiten auf. Die Ergebnisse beider Arbeiten deuten somit auf einen positiven Zusammenhang von Interaktionshäufigkeit und der KPL-Leistung hin.

Eine mögliche Erklärung des positiven Zusammenhangs von Interaktionshäufigkeit und KPL-Leistung ist, dass die Interaktionshäufigkeit ein Indikator für Personeneigenschaften wie Gründlichkeit, Ausdauer oder Motivation ist. Naumann (2015) fand, dass vor allem eine hohe Anzahl zielgerichteten Verhaltens positiv in komplexen Aufgaben wie dem Lesen von Hypertexten oder KPL ist. Auch Arbeit 3 zeigte, dass nicht nur die Quantität der Interaktion, sondern vor allem ihre Qualität eine wichtige Rolle spielt (vgl. Beckmann & Guthke, 1995; Greiff et al., 2018). Die qualitativen Unterschiede der Interaktionen wurden in Forschungsfrage 1c näher untersucht.

5.1.3. Die Rolle von Exploration (Forschungsfrage 1c)

Die Arbeiten 2 und 3 gingen der Frage nach der Rolle von Exploration beim KPL nach. In Arbeit 2 wurde der Zusammenhang von Exploration und Problemlöseleistung direkt untersucht. In Arbeit 3 kam ein explorativer Ansatz zur Anwendung, bei dem verschiedene Arten von Exploration differenziert betrachtet wurden. Die Ergebnisse beider Arbeiten deuten auf einen positiven Zusammenhang von lösungsrelevanter Exploration und KPL-Leistung hin, während in Arbeit 3 ein negativer Zusammenhang zwischen lösungsunabhängiger Exploration und KPL-Leistung vorlag.

Die Ergebnisse von Arbeit 2 zeigen, dass Exploration und KPL-Leistung positiv zusammenhängen. Dieses Ergebnis wurde trotz Variation der Effektstärke über sämtliche untersuchten Länder hinweg beobachtet. Ähnliches fanden Bell und Kozlowski (2008) sowie Dormann und Frese (1994). In Arbeit 3 wurde zwischen lösungsrelevanter und lösungsunabhängiger Exploration differenziert. Als lösungsrelevante Exploration wurde hier das wiederholte Ausführen zielgerichteter Interaktionen verstanden. Interaktionen, die für die Problemlösung nicht erforderlich waren, wurden hingegen als lösungsunabhängige Exploration verstanden. Die differenzierte Betrachtung dieser beiden Verhaltensweisen zeigte, dass lösungsunabhängige Exploration mit einer geringen KPL-Leistung einherging. Dieses Ergebnis schränkt die Ergebnisse von Arbeit 2 sowie von Dormann und Frese (1994) ein. Lösungsunabhängige Exploration scheint eher ein Ausdruck von Überforderung oder Ablenkung zu sein. Der Effekt der Exploration hängt daher vermutlich von deren Relevanz für die Problemlösung ab.

5.1.4. Komplexe Verhaltensmuster (Forschungsfrage 1d)

In Arbeit 3 wurden mit dem explorativen Ansatz der Sequenzanalyse komplexe Verhaltensmuster untersucht, die mit KPL-Leistung zusammenhängen. Es wurden verschiedene Verhaltensmuster gefunden, die sich größtenteils konsistent in beiden untersuchten KPL-Aufgaben zeigten.

Wie bereits im Kapitel 5.1.2 erwähnt, wurden bei beiden Aufgaben längere Verhaltenssequenzen bei richtigen Antworten beobachtet. Vergleicht man die gefundenen Verhaltensmuster hinsichtlich ihrer allgemeinen KPL-Leistung, so zeigten die

Schülerinnen und Schüler mit längeren Sequenzen auch hier höhere Leistung. Die höchste KPL-Leistung war bei Gruppen feststellbar, die lange Sequenzen überwiegend lösungsrelevanten Verhaltens aufwiesen. Diese Schülerinnen und Schüler schienen ihre Lösung mehrfach zu prüfen, bevor sie die Aufgabe abschlossen. Dies könnte Ausdruck von Personenmerkmalen wie Ausdauer oder Gewissenhaftigkeit bei der Aufgabenbearbeitung sein, die somit generell in komplexen Aufgaben förderlich sein könnten (Naumann, 2015). Jedoch führten lange Sequenzen lösungsunabhängigen Verhaltens nicht zum Erfolg. Schülerinnen und Schüler, die ein solches Verhalten zeigten, mögen zwar gewissenhaft sein, jedoch schaffen sie es nicht, relevante Informationen zu identifizieren.

Ein eher minimalistisches Vorgehen war ebenfalls erfolgreich, ging jedoch mit einer geringeren allgemeinen KPL-Leistung einher. Schülerinnen und Schüler in dieser Gruppe zeigten das minimal notwendige Verhalten, um das entsprechende Problem korrekt zu lösen. Da sie im Gegensatz zu der ausdauernden Gruppe ihre Ergebnisse nicht mehrfach prüften, unterliefen ihnen möglicherweise häufiger Flüchtigkeitsfehler, die unentdeckt blieben und somit nicht korrigiert wurden. Allerdings ist das minimalistische Vorgehen auch das effizientere Vorgehen, was bei wahrgenommenem Zeitdruck wiederum von Vorteil sein kann.

Sehr häufig wurde auch Rateverhalten beobachtet, wobei sich große Unterschiede in der Qualität des Ratens zeigten. In der Tickets-Aufgabe ging ein Drittel der korrekten Lösungen auf ein zielgerichtetes Rateverhalten zurück. Das heißt, statt mehrere Tickets zu vergleichen, um dann das günstigste auszuwählen, kauften die Schülerinnen und Schüler in dieser Gruppe das erste Ticket, das die Anforderungen der Aufgabe erfüllte. Mit dieser Heuristik hatten die Schülerinnen und Schüler somit nicht zwingend die optimale, aber eine ausreichende Lösung gefunden. Somit scheinen für diese Gruppe sowohl das Verständnis der Aufgabe als auch die Interaktion mit der Aufgabe kein Problem gewesen zu sein. Zielgerichtetes Raten könnte jedoch auf eine geringe Motivation hindeuten. Die allgemeine KPL-Leistung dieser Gruppe lag im mittleren Bereich.

Geringer war die KPL-Leistung von Schülerinnen und Schülern, die nicht zielgerichtet vorgingen, sondern eine scheinbar zufällige Lösung rieten oder die Aufgabe gar nicht lösten. Auch hier könnte eine geringe Motivation, aber auch grundlegende Schwierigkeiten beim KPL, eine Rolle spielen.

Eine mit der Gruppe des zielgerichteten Ratens vergleichbare KPL-Leistung konnte bei Schülerinnen und Schülern beobachtet werden, die unvollständiges zielgerichtetes Verhalten zeigten. Diese starteten vielversprechend, schafften es jedoch nicht das Problem zu lösen. Stattdessen wiederholten Schülerinnen und Schüler in dieser Gruppe oft ihr zielgerichtetes Verhalten, zeigten also lösungsrelevante Exploration, oder gingen zu lösungsunabhängiger Exploration über. Da diese Schülerinnen und Schüler keine extrem kurzen Sequenzen zeigten, kann davon ausgegangen werden, dass sie nicht besonders unmotiviert waren. Vielmehr schienen sie sich auf falsche Lösungen zu fokussieren.

5.2. Erklärung von Gruppenunterschieden durch Prozesse beim KPL (Forschungsfrage 2)

Um Forschungsfrage 2 zu beantworten wurden in Arbeit 2 Leistungsunterschiede beim KPL in Abhängigkeit vom Geschlecht und vom Migrationshintergrund sowie der Zusammenhang dieser Leistungsunterschiede mit Prozessmerkmalen untersucht. Die Ergebnisse bezüglich Forschungsfrage 2 werden in den folgenden Abschnitten dargestellt und diskutiert.

5.2.1. Unterschiede zwischen Mädchen und Jungen (Forschungsfrage 2a)

Analog zu den Ergebnissen der OECD (2014a) in Bezug auf Problemlösen im Allgemeinen zeigten die Ergebnisse von Arbeit 2 einen Leistungsunterschied beim KPL zugunsten von Jungen gegenüber Mädchen. Allerdings variierte dieser Unterschied zwischen den teilnehmenden Ländern und war in einigen Ländern nicht vorhanden oder sogar umgekehrt. Der beobachtete Leistungsunterschied zwischen Mädchen und Jungen ließ sich anhand der Problemlöseprozesse, Interaktions- und Explorationshäufigkeit statistisch erklären. Exploration konnte den Zusammenhang jedoch besser erklären als die Interaktionshäufigkeit. Somit scheint gerade die Exploration als Teilmenge der Interaktion einen entscheidenden Einfluss auf den Geschlechterunterschied beim KPL zu haben. Wurde für die Explorationshäufigkeit kontrolliert, verschwand der Geschlechterunterschied vollständig. Bei gleicher Explorationshäufigkeit zeigten Mädchen also eine ebenso gute KPL-Leistung wie Jungen. Eine mögliche Erklärung für die geringere Exploration der Mädchen ist, dass Mädchen durch ihre Sozialisation seltener zu explorativem Verhalten ermuntert werden. Cherney und London (2006) sowie Leaper und Friedman (2007) argumentieren, dass Mädchen durch eine geschlechtsspezifische Sozialisation, geschlechtsspezifische Spielsachen und die Konfrontation mit Stereotypen im Gegensatz zu Jungen selten zu explorativem Verhalten angeregt werden. Die OECD (2015a) berichtete außerdem, dass in den meisten Ländern Jungen früher der Zugang zu Computern gewährt wird und Mädchen weniger Freizeit mit Computern verbringen (OECD, 2015b). Dadurch könnten Mädchen eine geringere Selbstwirksamkeitserwartung im Umgang mit technischen Geräten haben, die wiederum zu passiverem Verhalten der Mädchen in technologiebasierten Aufgaben führen könnte (Naumann et al., 2014). Die unterschiedlichen Selbstwirksamkeitserwartungen von Mädchen und Jungen könnten somit ebenfalls eine Ursache der Leistungsunterschiede beim KPL sein.

5.2.2. Unterschiede zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund (Forschungsfrage 2b)

Die Ergebnisse von Arbeit 2 zeigen einen Leistungsunterschied beim KPL zugunsten von Schülerinnen und Schülern ohne Migrationshintergrund gegenüber denjenigen mit Migrationshintergrund. Auch dieser Leistungsunterschied zeigte sich bereits beim allgemeinen Problemlösen (OECD, 2014a). Dieser Leistungsunterschied ist jedoch ebenso vom jeweiligen Land abhängig. In einigen Ländern fand sich kein oder sogar ein umgekehrter Effekt des Migrationshintergrundes. Im Gegensatz zum Leistungsunterschied zwischen Mädchen und Jungen konnte der Unterschied zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund nicht durch die in Arbeit 2

untersuchten Problemlöseprozesse erklärt werden. Darüber hinaus zeigten sich in Arbeit 2 keine Verhaltensunterschiede abhängig vom Migrationshintergrund. Bei Schülerinnen und Schüler mit Migrationshintergrund konnte also eine geringere KPL-Leistung festgestellt werden, obwohl sie ein vergleichbares Verhalten in Bezug auf Interaktions- und Explorationshäufigkeit wie ihre Mitschülerinnen und Mitschüler ohne Migrationshintergrund an den Tag legten. Dies bestätigt die Ergebnisse von Sonnleitner et al. (2014), die darauf schließen, dass Schülerinnen und Schüler mit Migrationshintergrund keine Schwierigkeiten bei der Auswahl und Durchführung von Handlungen haben, wohl aber mit dem Transfer der gesammelten Informationen in deklaratives Wissen, also mit dem Verstehen und der Modellbildung (vgl. das Prozessmodell in Kapitel 2.2.2).

5.3. Zusammenfassende Diskussion

Ziel der vorliegenden Arbeit war es, die Wissensbasis über die Rolle von kognitiven Prozessen beim komplexen Problemlösen zu erweitern und zur Theoriebildung in diesem Bereich beizutragen. Durch die Integration bisheriger theoretischer Modelle konnte ein Modell entwickelt werden, das die Einordnung einzelner Prozesse beim KPL in den Kontext einer Intention stellt, wie von Funke (2003) vorgeschlagen. Dies erleichtert sowohl die Interpretation der Ergebnisse der drei Einzelarbeiten als auch die Einordnung bisheriger Forschungsergebnisse. Die drei vorgestellten Einzelarbeiten ergänzten sich hinsichtlich der Untersuchung unterschiedlicher kognitiver Prozesse beim KPL. Außerdem wurde untersucht, inwiefern diese Prozesse Leistungsunterschiede beim KPL zwischen Gruppen erklären können.

Die Ergebnisse zu Forschungsfrage 1a zeigen, dass die Aufgabenschwierigkeit für die Rolle kognitiver Prozesse eine wichtige Rolle spielen kann (vgl. Goldhammer et al., 2014). Darüber hinaus waren die Effekte der unterschiedlichen Planungsaspekte in Arbeit 1 in hohem Maße voneinander abhängig. Dies legt nahe, dass weitere Abhängigkeiten zwischen verschiedenen Prozessen bestehen könnten, die mit einer Analyse von komplexen Verhaltensmustern wie in Arbeit 3 aufgedeckt werden könnten. So zeigte sich in Arbeit 2 zwar ein positiver Effekt der Interaktionshäufigkeit, in Arbeit 3 konnte jedoch gezeigt werden, dass eine hohe Anzahl von Interaktionen keine hinreichende Bedingung für erfolgreiches KPL ist. Vielmehr kommt es auf die Qualität der Interaktion zwischen Person und Aufgabe an. Vor allem die Differenzierung zwischen lösungsrelevanter und lösungsunabhängiger Exploration war hier relevant.

Werden entscheidende Merkmale nicht differenziert betrachtet, kann dies zu fehlerhaften Schlüssen über die Rolle der jeweiligen Prozesse führen. Die Ergebnisse zu Forschungsfrage 1c zeigten, dass die Rolle von Exploration von ihrer Beziehung zur Aufgabenlösung abhängt. Während in Arbeit 2 der Effekt von Exploration insgesamt untersucht wurde, nimmt Arbeit 3 eine Differenzierung in lösungsrelevante und lösungsunabhängige Exploration vor. Dadurch konnte gezeigt werden, dass der in Arbeit 2 beobachtete positive Effekt der Exploration vermutlich nur durch die lösungsrelevante Exploration entstand. Die lösungsunabhängige Exploration hatte hingegen einen negativen Zusammenhang mit der KPL-Leistung. Betrachtet man lösungsunabhängige Exploration im Rahmen des intentionalen Handlungsmodells (vgl. Kapitel 2.2.2), so bietet das Modell hierfür eine Erklärung. Da Exploration eine Integration der gewonnenen Informationen in

ein mentales Modell nach sich zieht, wird dieses Modell durch die Exploration lösungsunabhängiger Informationen unnötig vergrößert. Dadurch verkomplizieren sich nachfolgend durchgeführte Planungs- und Reflexionsprozesse. Durch die Differenzierung der verschiedenen Arten der Exploration konnte die diesbezüglich zuvor bestehende Unklarheit der Ergebnisse aufgeklärt werden.

Das Erkenntnispotential, das in Logdaten liegt, ist noch lange nicht ausgeschöpft. Neue Analysemethoden ermöglichen immer vielfältigere und tiefere Einblicke in die Bearbeitungsprozesse bei computerbasierten Tests. Die vorliegende Arbeit impliziert einige Punkte, die bei der Analyse von Prozessen berücksichtigt werden sollten. So ist die Erstellung theoriegeleiteter Prozessmaße und die Verwendung von auf diese Theorien abgestimmten Methoden anzustreben. Außerdem sollten mögliche Interaktionseffekte sowohl zwischen verschiedenen Prozessen als auch zwischen den Prozessen und Aufgaben- oder Personenmerkmalen bedacht werden.

5.4. Limitationen

Bei den präsentierten Ergebnissen gibt es eine Reihe von Limitationen zu beachten. Wie bei jeder Forschung, die sich auf nicht beobachtbare kognitive Prozesse bezieht, muss beachtet werden, dass der Interpretation von Verhaltensmaßen im Sinne kognitiver Prozesse immer Annahmen zugrunde liegen, die fehlerbehaftet sein können (Wirth, 2004). Daher sollten diese Annahmen durch die Theorie bezüglich der untersuchten Prozesse gestützt werden.

Weiterhin sollte bei der Interpretation der Ergebnisse stets beachtet werden, dass es sich bei den durchgeführten Analysen um korrelative Zusammenhänge handelt. Eine eindeutige Zuweisung von Ursache und Wirkung ist in diesem Forschungsdesign nicht möglich. Die Interpretation der gefundenen Zusammenhänge kann zwar Hinweise auf mögliche ursächliche Zusammenhänge geben, es ist jedoch zwingend erforderlich, diese mit experimenteller Forschung zu überprüfen.

Eine weitere Limitation der Ergebnisse dieser Arbeit stellt die PISA-Stichprobe dar. Diese enthält ausschließlich Daten von 15-jährigen Schülerinnen und Schülern. Somit können die in dieser Arbeit berichteten Ergebnisse nicht ohne weiteres auf andere Personengruppen übertragen werden. Der Vorteil der Stichprobe ist hingegen die große Vielfalt an teilnehmenden Ländern und Kulturen. Somit können über Länder hinweg stabile Effekte zumindest als unabhängig von der Kultur betrachtet werden.

5.5. Fazit

Ziel dieser Arbeit war es, erfolgsrelevante Prozesse des komplexen Problemlösens zu identifizieren. Zudem sollte untersucht werden, ob sich Leistungsdisparitäten zwischen verschiedenen soziodemografischen Gruppen aufgrund dieser Prozessen erklären lassen. Dazu wurde ein intentionales Handlungsmodell der Prozesse beim KPL hergeleitet. Auf dieser Grundlage wurden die Zusammenhänge verschiedener Prozesse beim KPL mit der Leistung beim komplexen Problemlösen empirisch untersucht. Dadurch konnte die Wissensbasis in diesem Bereich erweitert und Annahmen über komplexe Problemlöseprozesse weiterentwickelt werden. Scheinbare Widersprüche bezüglich der

Rolle von Exploration konnten so aufgelöst werden. Außerdem ließen sich Leistungsunterschiede im KPL zwischen Mädchen und Jungen durch Bearbeitungsprozesse erklären.

Wie in Kapitel 1 beschrieben, handelt es sich bei KPL um eine Kompetenz, die gerade in einer sich ständig verändernden Welt an Bedeutung immer weiter zunimmt. Es ist anzunehmen, dass die beste Vorbereitung von Schülerinnen und Schülern auf die „Welt von morgen“ darin besteht, Kompetenzen zu entwickeln, die sie anpassungsfähig an sich wandelnde Herausforderungen machen – wie KPL (Trilling & Fadel, 2009). Daher leistet diese Arbeit einen Beitrag, um die Prozesse ebendieser Fähigkeit besser zu verstehen und somit die Grundlage für Konzepte zu schaffen, um junge Menschen beim Erwerb dieser Kompetenz zu unterstützen.

6. LITERATURVERZEICHNIS

Abbott, A. & Forrest, J. (1986). Optimal Matching Methods for Historical Sequences. *Journal of Interdisciplinary History, 14* (3), 471–494.

Albert, D. & Steinberg, L. (2011). Age differences in strategic planning as indexed by the tower of London. *Child Development, 82* (5), 1501–1517. https://doi.org/10.1111/j.1467-8624.2011.01613.x

Autor, D. H., Levy, F. & Murnane, R. J. (2003). The Skill Content of Recent Technological Change. An Empirical Exploration. *The Quarterly Journal of Economics, 118* (4), 1279–1333. https://doi.org/10.1162/003355303322552801

Beckmann, J. & Goode, N. (2017). Missing the Wood for the Wrong Trees. On the Difficulty of Defining the Complexity of Complex Problem Solving Scenarios. *Journal of Intelligence, 5* (2), 15. https://doi.org/10.3390/jintelligence5020015

Beckmann, J. & Guthke, J. (1995). Complex Problem Solving, Intelligence, and Learning Ability. In P. A. Frensch & J. Funke (Hrsg.), *Complex problem solving. The European perspective* (S. 177–200). Hillsdale, NJ: L. Erlbaum Associates.

Beckmann, J. F., Birney, D. P. & Goode, N. (2017). Beyond Psychometrics: The Difference between Difficult Problem Solving and Complex Problem Solving. *Frontiers in psychology, 8,* 1739. https://doi.org/10.3389/fpsyg.2017.01739

Bell, B. S. & Kozlowski, S. W. J. (2008). Active learning: effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology, 93* (2), 296–316. https://doi.org/10.1037/0021-9010.93.2.296

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M. et al. (2012). Defining Twenty-First Century Skills. In P. Griffin, B. McGaw & E. Care (Hrsg.), *Assessment and Teaching of 21st Century Skills* (S. 17–66). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-2324-5_2

Blech, C. & Funke, J. (2005, 24. August). *Dynamis review: An overview about applications of the Dynamis approach in cognitive psychology.* Deutsches Institut für Erwachsenenbildung. Zugriff am 11.11.2015. Verfügbar unter http://www.die-bonn.de/esprid/dokumente/doc-2005/blech05_01.pdf

Bos, W., Eickelmann, B. & Gerick, J. (2014). Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern der 8. Jahrgangsstufe in Deutschland im internationalen Vergleich. In W. Bos, B. Eickelmann, J. Gerick, F. Goldhammer, H. Schaumburg, K. Schwippert et al. (Hrsg.), *ICILS 2013. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern in der 8. Jahrgangsstufe im internationalen Vergleich* (S. 113–146). Münster: Waxmann.

Buchner, A. & Funke, J. (1993). Finite-state automata. Dynamic task environments in problem-solving research. *The Quarterly Journal of Experimental Psychology Section A, 46* (1), 83–118. https://doi.org/10.1080/14640749308401068

Cherney, I. D. & London, K. (2006). Gender-linked Differences in the Toys, Television Shows, Computer Games, and Outdoor Activities of 5- to 13-year-old Children. *Sex Roles, 54* (9-10), 717–726. https://doi.org/10.1007/s11199-006-9037-8

Cross, C. P., Copping, L. T. & Campbell, A. (2011). Sex differences in impulsivity: a meta-analysis. *Psychological bulletin, 137* (1), 97–130. https://doi.org/10.1037/a0021591

Dormann, T. & Frese, M. (1994). Error training. Replication and the function of exploratory behavior. *International Journal of Human-Computer Interaction, 6* (4), 365–372. https://doi.org/10.1080/10447319409526101

Dörner, D. (1989). Die Logik des Mißlingens. Strategisches Denken in komplexen Situationen (17. - 19. Tsd). Reinbek bei Hamburg: Rowohlt.

Dörner, D. & Funke, J. (2017). Complex Problem Solving: What It Is and What It Is Not. *Frontiers in psychology, 8,* 1153. https://doi.org/10.3389/fpsyg.2017.01153

Dörner, D., Kreuzig, H. W., Reither, F. & Stäudel, T. (Hrsg.). (1983). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität*. Bern: Huber.

Fischer, A. & Neubert, J. C. 2015. The multiple faces of complex problems: A model of problem solving competency and its implications for training and assessment (Heidelberg University Publishing, ed.). https://doi.org/10.11588/jddm.2015.1.23945

Frensch, P. A. & Funke, J. (1995). Definitions, Traditions, and a General Framework for Understanding Complex Problem Solving. In P. A. Frensch & J. Funke (Hrsg.), *Complex problem solving. The European perspective* (S. 3–22). Hillsdale, NJ: L. Erlbaum Associates.

Frey, C. B. & Osborne, M. (2013). *The Future of Employment*. University of Oxford, Oxford. Zugriff am 12.08.2019. Verfügbar unter http://sep4u.gr/wp-content/uploads/The_Future_of_Employment_ox_2013.pdf

Funke, J. (1988). Using Simulation to Study Complex Problem Solving. *Simulation & Games, 19* (3), 277–303.

Funke, J. (1992). *Wissen über dynamische Systeme: Erwerb, Repräsentation und Anwendung* (Lehr- und Forschungstexte Psychologie, Bd. 43). Berlin, Heidelberg: Springer.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning, 7* (1), 69–89. https://doi.org/10.1080/13546780042000046

Funke, J. (2003). *Problemlösendes Denken* (Einführungen und Allgemeine Psychologie, 1. Aufl.). Stuttgart: Kohlhammer. Verfügbar unter http://gbv.eblib.com/patron/FullRecord.aspx?p=1613642

Funke, J. (2010). Complex problem solving. A case for complex cognition? *Cognitive processing, 11* (2), 133–142. https://doi.org/10.1007/s10339-009-0345-0

Funke, J. & Greiff, S. (2017). Dynamic Problem Solving: Multiple-Item Testing Based on Minimally Complex Systems. In D. Leutner, J. Fleischer, J. Grünkorn & E. Klieme (Eds.), *Competence assessment in education. Research, models and instruments* (Methodology of Educational Measurement and Assessment, pp. 427–443). Cham: Springer International Publishing.

Gabadinho, A., Ritschard, G., Müller, N. S. & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software, 40* (4). https://doi.org/10.18637/jss.v040.i04

Goldhammer, F., Naumann, J. & Greiff, S. (2015). More is not Always Better: The Relation between Item Response and Item Response Time in Raven's Matrices. *Journal of Intelligence, 3* (1), 21–40. https://doi.org/10.3390/jintelligence3010021

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H. & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106* (3), 608–626. https://doi.org/10.1037/a0034716

Greiff, S. (2012a). Assessment and Theory in Complex Problem Solving - A Continuing Contradiction? *Journal of Educational and Developmental Psychology, 2* (1). https://doi.org/10.5539/jedp.v2n1p49

Greiff, S. (2012b). *Individualdiagnostik komplexer Problemlösefähigkeit* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 84). Zugl.: Heidelberg, Univ., Diss., 2010. Münster: Waxmann.

Greiff, S. & Funke, J. (2009). Measuring Complex Problem Solving: The MicroDYN approach. In F. Scheuermann & J. Björnsson (Hrsg.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (S. 157–163).

Greiff, S., Holt, D. V. & Funke, J. (2013). Perspectives on Problem Solving in Educational Assessment. Analytical, Interactive, and Collaborative Problem Solving. *The Journal of Problem Solving, 5* (2). https://doi.org/10.7771/1932-6246.1153

Greiff, S., Kretzschmar, A., Müller, J. C., Spinath, B. & Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *Journal of Educational Psychology, 106* (3), 666–680. https://doi.org/10.1037/a0035426

Greiff, S., Molnár, G., Martin, R., Zimmermann, J. & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments. A latent class approach. *Computers & Education, 126,* 248–263. https://doi.org/10.1016/j.compedu.2018.07.013

Greiff, S., Niepel, C., Scherer, R. & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61,* 36–46. https://doi.org/10.1016/j.chb.2016.02.095

Greiff, S., Wüstenberg, S. & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91,* 92–105.

He, Q. & von Davier, M. (2015). Identifying Feature Sequences from Process Data in Problem-Solving Items with N-Grams. In van der Ark,L. Andries, D. M. Bolt, W.-C. Wang, J. A. Douglas & S.-M. Chow (Hrsg.), *Quantitative Psychology Research* (Springer Proceedings in Mathematics & Statistics, Bd. 140, S. 173–190). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-19977-1_13

Keith, N. & Frese, M. (2005). Self-regulation in error management training. Emotion control and metacognition as mediators of performance effects. *Journal of Applied Psychology, 90* (4), 677–691. https://doi.org/10.1037/0021-9010.90.4.677

Koeppen, K., Hartig, J., Klieme, E. & Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. *Zeitschrift für Psychologie / Journal of Psychology, 216* (2), 61–73. https://doi.org/10.1027/0044-3409.216.2.61

Leaper, C. & Friedman, C. K. (2007). The Socialization of Gender. In J. E. Grusec & P. D. Hastings (Hrsg.), *Handbook of Socialization: Theory and Research* (S. 561–587). New York: Guilford Publications.

Leutner, D., Funke, J., Klieme, E. & Wirth, J. (2005). Problemlösefähigkeit als fächerübergreifende Kompetenz. In E. Klieme, D. Leutner & J. Wirth (Hrsg.), *Problemlösekompetenz von Schülerinnen und Schülern. Diagnostische Ansätze, theoretische Grundlagen und empirische Befunde der deutschen PISA-2000-Studie* (1. Aufl, S. 11–19). Wiesbaden: VS Verl. für Sozialwiss.

Mayer, R. E. & Wittrock, M. C. (2006). Problem Solving. In P. A. Alexander & P. H. Winne (Hrsg.), *Handbook of educational psychology* (2nd ed, S. 287–304). Mahwah, N.J: Erlbaum.

Naumann, J. (2015). A model of online reading engagement. Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior, 53,* 263–277. https://doi.org/10.1016/j.chb.2015.06.051

Naumann, J., Goldhammer, F., Rölke, H. & Stelter, A. (2014). Erfolgreiches Problemlösen in technologiebasierten Umgebungen: Wechselwirkungen zwischen Interaktionsschritten und Aufgabenanforderungen. *Zeitschrift für Pädagogische Psychologie, 28* (4), 193–203. https://doi.org/10.1024/1010-0652/a000134

Neubauer, A. C. (1990). Speed of information processing in the Hick paradigm and response latencies in a psychometric intelligence test. *Personality and Individual Differences, 11* (2), 147–152.

Neubert, J. C., Kretzschmar, A., Wüstenberg, S. & Greiff, S. (2015). Extending the Assessment of Complex Problem Solving to Finite State Automata. *European Journal of Psychological Assessment, 31* (3), 181–194. https://doi.org/10.1027/1015-5759/a000224

OECD. (2013). PISA 2012 assessment and analytical framework. Mathematics, reading, science, problem solving and financial literacy (PISA). Paris: OECD.

OECD. (2014a). *PISA 2012 Results: Creative Problem Solving. Students' skills in tackling real-life problems* (Volume V). Paris: OECD Publishing. Verfügbar unter http://dx.doi.org/10.1787/9789264208070-en

OECD. (2014b). *PISA 2012. Technical Report*. Paris: OECD Publishing. Zugriff am 14.10.2019. Verfügbar unter https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD. (2015a). Students, computers and learning. Making the connection. Paris: OECD Publishing.

OECD. (2015b). *The ABC of Gender Equality in Education. Aptitude, Behaviour, Confidence* (PISA). Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264229945-en

Perez, S., Massey-Allard, J., Butler, D., Ives, J., Bonn, D., Yee, N. et al. (2017). Identifying Productive Inquiry in Virtual Labs Using Sequence Mining. In E. André, R. Baker, X. Hu, M. M. T. Rodrigo & B. Du Boulay (Hrsg.), *Artificial Intelligence in Education* (Bd. 10331, S. 287–298). Cham: Springer International Publishing.

Robitzsch, A., Kiefer, T. & Wu, M. (2019). TAM: Test Analysis Modules (Version R package version 3.3-10) [Computer software]. Verfügbar unter https://CRAN.R-project.org/package=TAM

Sonnleitner, P., Brunner, M., Keller, U. & Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *Journal of Educational Psychology, 106* (3), 681–695. https://doi.org/10.1037/a0035506

Stadler, M., Fischer, F. & Greiff, S. (2019). Taking a Closer Look. An Exploratory Analysis of Successful and Unsuccessful Strategy Use in Complex Problems. *Frontiers in psychology, 10,* 248. https://doi.org/10.3389/fpsyg.2019.00777

Stadler, M., Niepel, C. & Greiff, S. (2019). Differentiating between static and complex problems. A theoretical framework and its empirical validation. *Intelligence, 72,* 1–12. https://doi.org/10.1016/j.intell.2018.11.003

Stemmann, J. & Lang, M. (2018). Eignet sich die logfilegenerierte Explorationsvollständigkeit als Prozessindikator für den Wissenserwerb im problemlösenden Umgang mit technischen Alltagsgeräten? *Journal of Technical Education, 6* (1), 185–199.

Studer, M. (2013). WeightedCluster Library Manual. A practical guide to creating typologies of trajectories in the social sciences with R. https://doi.org/10.12682/lives.2296-1658.2013.24

Trilling, B. & Fadel, C. (2009). *21st century skills. Learning for life in our times*. San Francisco, Calif.: Jossey-Bass a Wiley Imprint.

Unterrainer, J. M. & Owen, A. M. (2006). Planning and problem solving: from neuropsychology to functional neuroimaging. *Journal of physiology, Paris, 99* (4-6), 308–317. https://doi.org/10.1016/j.jphysparis.2006.03.014

Wirth, J. & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme. In E. Klieme, D. Leutner & J. Wirth (Hrsg.), *Problemlösekompetenz von Schülerinnen und Schülern. Diagnostische Ansätze, theoretische Grundlagen und empirische Befunde der deutschen PISA-2000-Studie* (1. Aufl, S. 55–72). Wiesbaden: VS Verl. für Sozialwiss. Zugriff am 19.08.2019. Verfügbar unter https://archiv.ub.uni-heidelberg.de/volltextserver/8248/1/WirthFunke_2005_HFA.pdf

Wirth, J. (2004). *Selbstregulation von Lernprozessen* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 39). Zugl.: Berlin, Humboldt-Univ., Diss., 2003. Münster: Waxmann.

Wittmann, W. W. & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science, 21* (4), 393–409. https://doi.org/10.1002/sres.653

Wüstenberg, S., Greiff, S., Molnár, G. & Funke, J. (2014). Cross-national gender differences in complex problem solving and their determinants. *Learning and Individual Differences, 29,* 18–29. https://doi.org/10.1016/j.lindif.2013.10.006

ANHANGVERZEICHNIS

ANHANG A: ARBEIT 1

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education, 128*, 1–12. https://doi.org/10.1016/j.compedu.2018.08.004

Computers & Education 128 (2019) 1–12

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

The role of planning in complex problem solving

Beate Eichmann[a,*], Frank Goldhammer[a], Samuel Greiff[b], Liene Pucite[c], Johannes Naumann[d]

[a] German Institute for International Educational Research, Centre for International Student Assessment, Schloßstraße 29, 60486, Frankfurt am Main, Germany
[b] University of Luxembourg, 2, avenue de l'Université, L-4365, Esch-sur-Alzette, Luxembourg
[c] Goethe University Frankfurt, Theodor-W.-Adorno-Platz 1, 60323, Frankfurt am Main, Germany
[d] University of Wuppertal, Gauss Straße 20, 42119, Wuppertal, Germany

ARTICLE INFO

Keywords:
Complex problem solving
Planning
Computer-based assessment
Log data
PISA

ABSTRACT

Complex problem solving (CPS) is a highly transversal competence needed in educational and vocational settings as well as everyday life. The assessment of CPS is often computer-based, and therefore provides data regarding not only the outcome but also the process of CPS. However, research addressing this issue is scarce. In this article we investigated planning activities in the process of complex problem solving. We operationalized planning through three behavioral measures indicating the duration of the longest planning interval, the delay of the longest planning interval and the variance of intervals between each two successive interactions. We found a significant negative average effect for our delay indicator, indicating that early planning in CPS is more beneficial. However, we also found effects depending on task and interaction effects for all three indicators, suggesting that the effects of different planning behaviors on CPS are highly intertwined.

1. Introduction

1.1. Theoretical background

The ability to solve complex problems is an essential competence in both education and everyday life, and is required for active participation in today's society. Ongoing globalization and digitalization confront people with an increasingly complex environment that demands numerous problems to be solved in personal life as well as at the workplace (Fischer, Greiff, & Funke, 2012). Furthermore, problem solving is the basis of many scholastic learning processes and is therefore regarded as a fundamental goal of education (OECD, 2013). From this perspective, the question stands which kind of behavioral engagement with a problem solving task predicts successful task completion. This study investigates the relationship between planning behavior and performance in complex problem solving (CPS) using computer-generated log data from a large-scale assessment.

According to Mayer and Wittrock (2006) a problem occurs when someone wants to achieve a goal and no obvious solution is available. This broad definition refers to many academic and real-world tasks in the field of problem solving. In this paper, we focus on CPS, which is frequently needed in real life. In contrast to non-complex problems, in CPS the given state, the goal state and the barriers between these states are complex, i.e. they can change dynamically and are opaque (not all information is presented at the

* Corresponding author.
E-mail addresses: beate.eichmann@dipf.de (B. Eichmann), goldhammer@dipf.de (F. Goldhammer), samuel.greiff@uni.lu (S. Greiff), pucite@em.uni-frankfurt.de (L. Pucite), j.naumann@uni-wuppertal.de (J. Naumann).

https://doi.org/10.1016/j.compedu.2018.08.004
Received 14 December 2017; Received in revised form 31 July 2018; Accepted 4 August 2018
Available online 06 August 2018

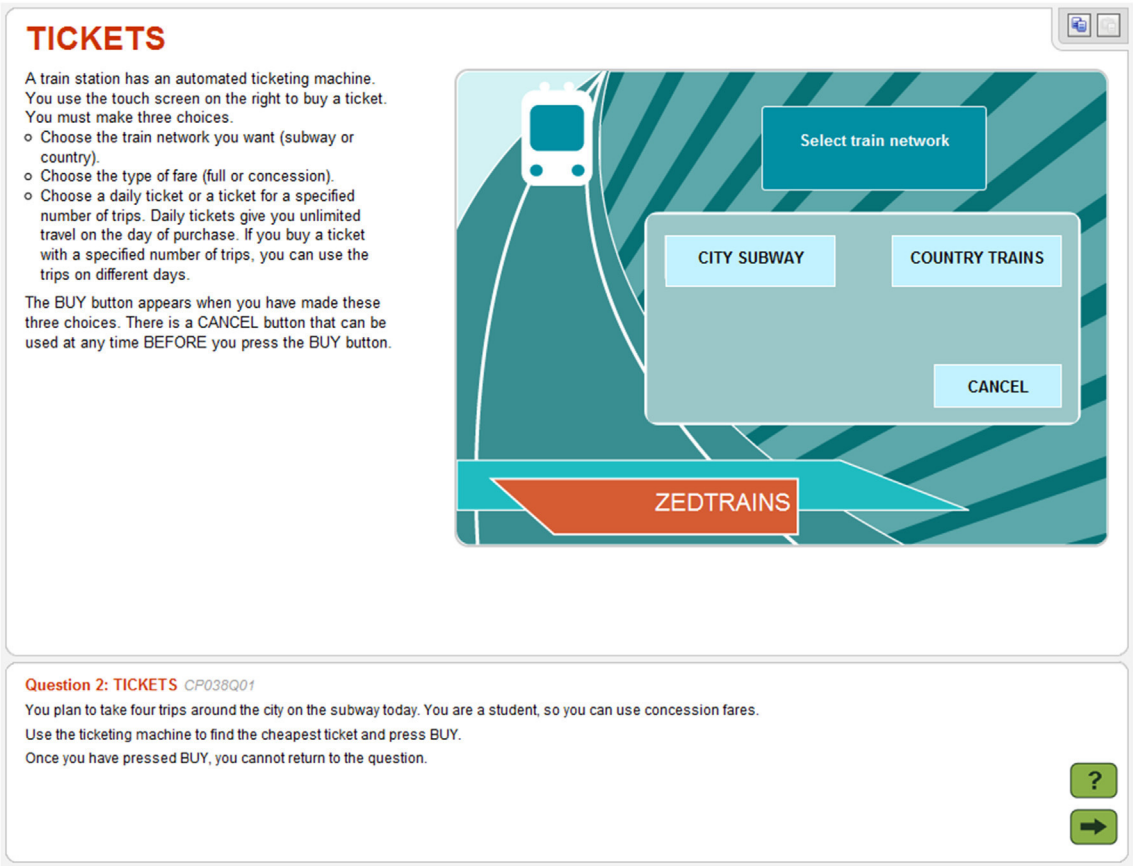

Fig. 1. CPS-task from PISA 2012.

outset) (Frensch & Funke, 1995). Hence, the problem solver has to interact with the problem to overcome these barriers and solve the problem. In the scientific literature many similar definitions of CPS exist, which all stress the importance of dynamics, opacity, and interactivity. To account for the interactivity of complex problems, CPS is mostly assessed via computer-based tasks, as it was the case in the Programme for International Student Assessment (PISA) 2012 (OECD, 2013). PISA measures curricular and cross-curricular competencies of fifteen-year-olds in the participating countries every three years. An example of a problem solving task, which is taken from the PISA 2012 problem solving assessment, is shown in Fig. 1. It displays a simulated ticket machine and an instruction to buy a specific ticket. This scenario can be regarded as a problem: there is a given state (not having a ticket), a goal state (having a ticket) and barriers between them (the user interface of the ticket machine). The problem is complex, in that not all information is present at the outset (e.g. it is not shown how different fare types are selected) and it therefore requires a number of interactions to gain the information needed to solve the task. Because CPS is considered a cross-curricular domain, its assessment can take many different forms and the tasks used in PISA 2012 are rather heterogeneous in nature (Buchner, 1995; Leutner, Funke, Klieme, & Wirth, 2005).

Several cognitive processes are involved in the solving of complex problems. The OECD (2013) established a framework of problem solving that is based upon the work of several cognitive psychologists (Baxter & Glaser, 1997; Blech & Funke, 2005, 2010; Bransford, Brown, & Cockling, 1999; Funke & Frensch, 2007; Greiff & Funke, 2008; Klieme, 2004; Mayer & Wittrock, 2006; Osman, 2010; Reeff, Zabal, & Blech. C., 2006; Vosniadou & Ortony, 1989; Wirth & Klieme, 2003). According to this framework, four main processes are involved in problem solving: “exploring and understanding”, “representing and formulating”, “planning and executing”, and “monitoring and reflecting”. The process of “exploring and understanding” involves interaction with the problem environment to gain information. Furthermore, the given and the found information should be understood by creating mental models of the pieces of information. The process of “representing and formulating” covers the creation of a mental representation of the problem situation as a whole. This includes the selection and integration of information, and the formulation of hypotheses about the problem. “Planning and executing” involve the setting of goals and sub-goals and the selection and execution of steps to achieve these goals. The last process, “monitoring and reflecting”, requires monitoring the progress towards the goals and, if necessary, for the individual to adjust their behavior as well as reflecting on assumptions and solutions. However, not all of the processes have to be present in each particular problem solving process. Furthermore, according to Lesh and Zawojewski (2007), the human brain is capable of parallel information processing, so it is also assumed that in CPS several processes can happen simultaneously. Unterrainer and Owen (2006) argue that even in non-complex problems people tend to switch between the different processes of problem solving.

Notably, planning occurs in three out of the four processes in the PISA framework (OECD, 2013): “exploring and understanding”, “representing and formulating” and “planning and executing” all refer to planning as it is defined by Unterrainer and Owen (2006). They argue that planning requires an individual to form a representation of the current state and goal state of a problem, and also to conceive of a sequence of actions, which transform the current state into the goal state. In the PISA framework, understanding the problem and creating a representation of the problem could therefore be understood as forming a descriptive representation of the possible states of the problem and actions to switch between them, while what is explicitly called “planning” in the framework should target the process of selecting the specific actions required to reach the desired goal (cf. Novick & Bassok, 2005). Formulating hypotheses about the problem refers to both aforementioned parts of planning, since the problem solver is required to formulate hypotheses about both the structure of the problem and steps to solve it. This shows that planning is an important part of CPS, which serves several purposes. In this article we follow this understanding of planning and investigate empirically the relevance of planning in CPS. Although planning is part of many theories about CPS (as in the PISA framework), empirical studies addressing this process and its relevance for success in CPS are scarce. In the following, we summarize relevant results from this domain.

1.2. Literature review

Like any cognitive process, planning cannot be observed directly. Thus, the occurrence of planning has to be inferred from behavior. To the advantage of the researcher, CPS is an interactive process, which makes it empirically accessible for detailed investigation (Greiff, Niepel, Scherer, & Martin, 2016). Although several studies investigated behavior while solving complex and non-complex problems, there are no studies addressing planning in CPS. Therefore, we examine both research about planning in related domains and studies addressing behavior in CPS that is related to planning: Albert and Steinberg (2011) found that first move latency is positively related to performance in non-complex problem solving tasks. They argue that individuals with higher first move latency complete their initial planning phase and therefore show a higher performance rate, in comparison to people with lower first move latency. Furthermore, when instructed to plan ahead the entire process of solving the task before starting to interact with it, the performance in problem solving was improved (Unterrainer & Owen, 2006). Both of these studies demonstrate the importance of prior planning in problem solving. However, both studies also used the “Tower of London” task for their investigations, which is a non-complex problem solving task according to the definition by Frensch and Funke (1995), so it remains unclear whether these findings hold for complex problems. As already mentioned, in complex problems the situation can change dynamically and not all of the information required to find a solution is presented at the outset, so in this case it is not possible to plan ahead completely before having interacted with the problem. Along this line of thinking, Greiff et al. (2016) found a negative correlation between performance in CPS and the frequency of interactions with the task, indicating that fewer and therefore presumably better planned actions also lead to higher performance in complex problems. Although this result indicates that it is beneficial to invest time in cognitive activities such as planning, instead of permanently interacting with a complex problem task, it does not show how this time should be distributed over the process of CPS. Another process measure, which might be related to time spent on planning, is the overall time spent on a task. Goldhammer et al. (2014) argue that, according to dual processing theory, the interpretation of time spent on a task depends on the nature of the task and the cognitive processes required to solve the task. In tasks that require non-routine behavior (like CPS tasks), time spent on a task can be regarded as an indicator of effort, so it can be assumed that in these scenarios students actually spend their time working on the task. In contrast, in tasks which only require routine behavior, time spent on a task can be seen as an indicator of having trouble solving the task. Supporting this theory, Goldhammer et al. (2014) showed that in problem solving tasks a higher amount of time spent on a task was associated with higher performance. Furthermore, they found that the effect of time spent depends on task difficulty and being strong especially in hard tasks. These results also indicate that investing time in cognitive activities (such as planning) might be beneficial in CPS. Also in line with this notion, Naumann and Goldhammer (2017) found that the investment of time in digital reading, a domain which can also be conceived as solving information problems (Brand-Gruwel, Wopereis, & Walraven, 2009; Rouet & Le Bigot, 2007) was positively predictive of successful task completion especially in hard tasks. Moreover, these authors showed that positive effects of time were accelerated with task's navigation demands, thus with the problem space becoming larger, and more opaque.

However, the relationship between time spent on a task and performance in CPS does not seem to be purely linear. Greiff et al. (2016) found a quadric relation between the time spent on a task and performance, where times that were too long or too short were associated with poor performance. Likewise, Naumann and Goldhammer (2017) found a negatively sloped quadratic trend for the time on task relation to performance in digital reading. However, where tasks had high navigation demands, this quadratic trend was more linear, possibly indicating that in these tasks positive effects of cognitively demanding activities such as planning levelled off not that early. In other studies, task-dependent effects of process measures were also found (e.g. Goldhammer, Naumann, & Greiff, 2015). For example, Naumann, Goldhammer, Rölke, and Stelter (2014) found that in technology-based problem solving the effect of the number of performed interactions on performance was moderated by the minimum number of interactions the task required. Hence, in a broad domain such as CPS it can be assumed that the effects of process measures on performance depend on certain task characteristics such as difficulty or the number of required interactions.

1.3. The current study

Although planning is an important part of many theories about CPS (e.g. Dörner, 1989; Leutner et al., 2005; Mayer & Wittrock, 2006) there is no research about the impact of planning on performance in CPS (to the best of our knowledge). Furthermore, there is no research that we are aware of that addresses the effects of different aspects of planning. In addition to the recent findings that

general indicators of planning have a positive impact on performance in CPS, we attempt to inspect several aspects of planning by using different process measures: we investigate whether the length of planning intervals, the time when planning takes place, and the variation of planning time over the whole CPS process influence the performance in CPS, and therefore include three different aspects of planning during CPS tasks. For each of those three aspects we developed a process measure that addresses the timespan between two successive interactions. Our definition of planning refers to understanding, representing, formulating, and planning from the PISA framework (OECD, 2013). While exploring and executing explicitly refer to visible action, this leaves the processes that represent planning, as well as monitoring and reflecting to happen in the time between. Therefore, process indicators based on the time between interactions should capture planning behavior while also correlating with monitoring and reflecting and task-irrelevant behavior.

We will further introduce the process indicators we used for planning in the following paragraphs. Because research in this area is scarce, we concentrated on the following explorative research questions:

Research question 1: Is the length of planning intervals related to performance in CPS?

According to several studies, planning activities have a positive impact on problem solving. For example, Albert and Steinberg (2011) showed in their study that first-move latency as a predictor of planning behavior is positively related to task outcome in problem solving. We extend this measure and not only take into account the time before the first move, but also include planning time that appears at any point in the problem solving process. Due to the opacity of complex problems and the need for exploration, we argue that planning might occur at a later point in time, after the collection of necessary information, and not just and exclusively at the beginning.

We use the duration of the longest interval, in which no interaction with the problem occurs, as an indicator of the length of planning intervals (duration indicator). In the example task (Fig. 1), this would be the longest time without a button being pressed. Obviously, this measure does not cover all the planning that happens during the process but it gives information about the longest instance when planning activities might take place. In doing so, we are able to cover main planning activities no matter if they appear at the beginning of the process or at a later stage. However, it is unclear whether planning or any other activities take place during this time. Nevertheless, the results of Goldhammer et al. (2014) indicate that in problem solving tasks it is beneficial to invest more time in the task, so we assume that time spent on a task is mostly spent by engaging in task-relevant cognitive activities such as planning.

We further investigate the relationship between the duration indicator and performance by testing whether the effect varies between tasks. The results of several studies show that the effects of process measures in problem solving are dependent on task characteristics, so this might also be true for planning indicators (e.g. Goldhammer et al., 2014; Goldhammer et al., 2015; Naumann & Goldhammer, 2017; Naumann et al., 2014). The tasks used in this study vary in many respects, such as in level of difficulty, complexity, or in the domain of knowledge, and cover a wide range of everyday problems. Therefore different tasks might require different styles of planning. For example, tasks that require more interactions to get to a solution might also require the individual to plan for longer.

Research question 2: Is the time when planning takes place related to performance in CPS?

The second aspect of planning behavior that we investigate is the time when planning takes place in CPS. As mentioned before, Albert and Steinberg (2011) showed that planning at the beginning of a non-complex problem solving process is beneficial, while there is no research addressing the issue of when planning in CPS is beneficial. In the PISA framework (OECD, 2013) it is stated that planning can happen at any time, but it leaves open when or to what extent planning is beneficial for the task outcome. In complex problems it might not be optimal to make a plan at the beginning of the process but rather after some exploration; or there might not be an optimal time for planning at all.

To investigate this question, we use the delay of the longest interval, in which no interactions take place, as a measure for the time when (most of) the planning might take place. We use the delay of this interval from the beginning of the process as our indicator for research question 2 (delay indicator). In the example task (Fig. 1) this would be the time until the longest pause (when no buttons are pressed) occurs in the process.

In a manner similar to what we assumed in research question 1, the effect of the delay indicator could also depend on task characteristics. Thus, we investigate when planning should optimally occur and whether the effect varies between tasks. The heterogeneity of CPS tasks leads to many different dimensions, in which the tasks may vary. For example, a task that has a very rich environment that can take many different states and requires many interactions to find the solution might also require more exploration before meaningful planning can take place. Conversely, in less rich environments planning could occur earlier.

Research question 3: Is the variation of planning time related to performance in CPS?

Another aspect of planning that might influence the performance of the individual in CPS is the variation of planning time. Variation of planning time over several phases of similar length means that the interactions are performed by the problem solver at an overall constant speed. There are neither especially long nor especially short intervals between the interactions, so we assume that such variation of planning time indicates similar efforts in planning before every interaction. If, in contrast, people systematically allocate the time they take for planning their actions when necessary, this should cause larger variance in planning time across the

course of completing a task. Because exploration is a central feature in CPS, we assume that a higher variation in planning time is beneficial: both phases of high exploration and phases of planning are necessary to solve a task.

We use the variance in the length of the intervals between two successive interactions as a measure of variation of planning time (variance indicator). With this measure, we take into account the whole process of CPS. Therefore, it is an extension of the two indicators introduced in the previous research questions, since these indicators only consider the longest planning interval.

The predictive power of the variance indicator, once again, might depend on task characteristics: for example, in a task that requires many similar interactions (like pressing the same button several times) it might be beneficial to perform only a few longer planning intervals when planning a sequence of actions, while in a task that requires different interactions, it might be beneficial to plan in more homogenous intervals when planning one action at a time. For this reason we investigate whether the effect of the variance indicator varies across tasks.

Research question 4: Are there interaction effects between the three process indicators mentioned above?

Few studies consider several behavioral indicators in CPS at once. We investigate interaction effects between the three indicators mentioned above, thus endorsing a comprehensive view on different aspects of planning. As a result, we wish to better understand the relationship between planning behavior and performance, and investigate whether the expected effects of our three indicators depend on each other. For example, the expected effect of the duration indicator could depend on the delay indicator, since a late planning phase might not be as crucial as an early planning phase. On the other hand, the effect of the duration indicator could also depend on the variance indicator: the duration of the longest phase might only play a role if there is any variation between the phases of planning.

2. Material and methods

2.1. Sample

For the analyses reported here, data from the fifth cycle of the PISA study (PISA, 2012) were used. In the present study, only data from the CPS tasks of German students (N = 1350) were used. After clearing invalid data, which might have occurred due to technical issues during data collection, N = 1346 students were left in the sample, of which 48.7% were female.

2.2. Instruments

In PISA 2012, problem solving was assessed via 42 computer-based tasks, which were organized into 16 units of two or three tasks each. An example task is shown in Fig. 1. Tasks in the same unit shared common stimulus material with only minor deviations. Students completed one or two (out of four) different problem solving clusters, which consisted of four units (ten to eleven tasks) each. The order of tasks within units and the order of units within clusters was always the same. Students were not able to return to a former task after finishing it.

According to OECD (2013), only very basic ICT skills were required to work on the computer-based problem solving tasks. To cover a wide range of difficulty, the tasks assessing problem solving varied across the following characteristics: amount, representation, and disclosure of information, internal complexity, distance to the goal, degree of abstraction, familiarity of the context, and reasoning skills required. For a detailed explanation of these task characteristics see OECD (2013). Notably, the tasks required interactive behavior like clicking on virtual buttons and sliders, dragging and dropping, operating simulated machines, exploring simulated environments, and manipulating variables. The response formats used included simple and complex multiple-choice tasks that were answered by clicking radio buttons, tasks that required shapes to be selected and dragged into position, tasks in which selections had to be made from pull-down menus, tasks that required parts of diagrams to be drawn or highlighted, and text boxes (OECD, 2013).

2.3. Scoring

The correctness of students' answers was derived from the computer-generated log data. For overall CPS performance, the response coding from OECD (2015) was used: the responses were coded either as correct (1) or incorrect (0), where partially solved tasks were coded as incorrect. We did not use the original scoring rules because they included partial credit based on students' behavior during the task. However, we were only interested in the effect of students' behavior on task outcome. Out of the total of 42 tasks, only 28 are considered to be interactive and therefore meet the definition of complex problems (Frensch & Funke, 1995). Only these tasks were used for data analyses. Two more tasks had to be excluded because the response included free text input, which was not recorded in the log data, so the correctness of the response could not be inferred. Another two tasks were excluded due to the respective log files including undefined events (clicks on elements with ambiguous IDs). For another three tasks, there was no valid log data in the German sample due to invalid and missing data. Therefore, after selecting data from only German students and clearing invalid data, a total of 21 tasks were left, out of the 28 tasks that could be considered complex (for task examples see Fig. 1 or OECD, 2013).

In accordance with our research questions, the three process indicators mentioned above were extracted from the log data: for research question 1 we used the duration of the longest interval between two successive interactions (duration indicator); for research

Table 1
Descriptive statistics of the three mean-centered indicators.

Indicators	SD	min	max	Skewness
Duration indicator	0.79	−2.14	2.69	0.16
Delay indicator	1.21	−0.59	4.48	1.97
Variance indicator	0.08	−0.04	1.31	4.71

question 2 we used the delay of the longest interval between two successive interactions from the beginning of the process (delay indicator); and for research question 3 we used the variance of times between each two successive interactions (variance indicator). The indicators were extracted for each combination of person and task for which data was available depending on the booklet design.

2.4. Procedure

The computer-based assessment of problem solving was part of the PISA 2012 study, which covered the domains of mathematics, reading, science, problem solving, and financial literacy. The computer-based assessment included the domains of problem solving, mathematics and reading. Students were administered two computer-based task clusters of which none, one, or both were problem solving clusters. An example of a computer-based problem-solving task is shown in Fig. 1. The computer-based assessment took place after the completion of the paper-based PISA tasks. Students were given 20 min per cluster.

2.5. Data preparation

The three indicators were inspected for outliers. An outlier was defined as a data point three standard deviations above/below the mean of the corresponding task, as suggested by Goldhammer et al. (2015). Overall, the sample comprised 6.63% outliers, which were replaced by the value at three standard deviations above/below the respective average as suggested by Goldhammer et al. (2014). Because the distributions of the three indicators showed pronounced skewness (duration indicator: 1.97, delay indicator: 4.99, variance indicator: 7.95), we normalized the distributions, using the logarithms of the three indicators. Since the ranges of the delay indicator and the variance indicator included zero, +1 was added to enable the logarithmic transformation. After this, all three indicators were centered around their respective grand mean. Table 1 shows the descriptive statistics of the three indicators after the described data preparation.

2.6. Data analyses

We estimated generalized linear mixed models (GLMM) to investigate our research questions using the package lme4 in the R-environment (Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2016). The GLMM framework allows modeling the probability of a correct response as a function of fixed and random effects (Baayen, Davidson, & Bates, 2008). In this way, we could model the overall (fixed) effects and the task-dependent (random) effects together, and at the same time, take into account the multilevel structure of our data (students nested in schools). For research questions 1 to 3, we estimated two models for each of our three indicators respectively: one model only including the overall (fixed) effect and one model containing both the overall and the task-dependent random effects. For the duration indicator, both models are shown in equation (1) and equation (2):

$$\eta_{pi} = \beta_0 + b_{0i} + b_{0p} + b_{0s} + \beta_1 M_{pi} \tag{1}$$

$$\eta_{pi} = \beta_0 + b_{0i} + b_{0p} + b_{0s} + (\beta_1 + b_{1i}) M_{pi} \tag{2}$$

In this model η_{pi} denotes the logit of the probability of a successful solution for person p completing task i, β the fixed effects and b the random effects. M represents the duration indicator. In the model, persons are nested within schools s.

The models for the other two indicators look the same, with the duration indicator being replaced by the delay, and variance indicator, respectively. Subsequently, we compared the two models for each indicator to test the task-dependent effects for significance. We used the likelihood ratio test and the Akaike Information Criterion (AIC) as criteria for model fit (Baayen et al., 2008; De Boeck et al., 2011). From the respective models, that fitted the data best, we took the estimates for the overall effects of our three process indicators as well as their p-values. For research question 4, we did not inspect random effects of interactions between indicators, since these effects could hardly be interpreted. Therefore, we estimated a model including only overall (fixed) effects of our three indicators and interaction effects between them. This model is shown in equation (3):

$$\eta_{pi} = \beta_0 + b_{0i} + b_{0p} + b_{0s} + \beta_1 M_{pi} + \beta_2 D_{pi} + \beta_3 V_{pi} + \beta_4 M_{pi} D_{pi} + \beta_5 M_{pi} V_{pi} + \beta_6 D_{pi} V_{pi} + \beta_7 M_{pi} D_{pi} V_{pi} \tag{3}$$

In this model, M, D and V represent the duration, delay and variance indicator respectively.

3. Results

The model comparisons are shown in Table 2. In the table, each comparison starts with ModelX.1 representing models including

Table 2
Model comparison; K is the number of parameters in the respective model, ΔK is the difference in the number of parameters.

Indicator	Model	K	AIC	BIC	logLik	deviance	X^2	ΔK	p	
Duration	Model1.1	5	9733.5	9769.0	−4861.8	9723.5	46.61	2	< 0.001	***
	Model1.2	7	9690.9	9740.6	−4838.4	9676.9				
Delay	Model2.1	5	9727.2	9762.6	−4858.6	9717.2	16.20	2	< 0.001	***
	Model2.2	7	9715.0	9764.6	−4850.5	9701.0				
Variance	Model3.1	5	9734.8	9770.3	−4862.4	9724.8	34.26	2	< 0.001	***
	Model3.2	7	9704.5	9754.2	−4845.3	9690.5				

Note. *p < 0.05, **p < 0.01, ***p < 0.001.

only overall (fixed) effects and proceeds with ModelX.2 representing models including both overall (fixed) and task-dependent random effects. The model comparisons for research questions 1–3 show that the models including both overall (fixed) and task-dependent random effects fit the data better than models including only overall effects, as shown in Table 2. In all model comparisons the more complex model has the better fit according to the likelihood ratio test and AIC. Therefore we used the estimates of those models to investigate the overall and task-dependent effects of our three indicators.

The model parameters of Model1.2, Model2.2 and Model3.2 are shown in Table 3. For the task-dependent random effects, the variance and standard deviation as well as the correlation between the random effect of the indicator and the task intercept are presented. For the fixed effects, estimates, standard errors, z-values, and p-values are shown.

The task-dependent random effects are depicted in Fig. 2. The effects of the three indicators for every single task are shown as a function of task easiness.

Research question 1: For the duration of the longest interval no significant fixed effect on the probability of success was found. However, a significant random effect was found for this indicator. The correlation between task easiness and the random effect of the indicator was $r = -0.04$. The random effect shows that the duration of the longest interval varies between tasks but is not related to task difficulty.

Research question 2: For the delay indicator, a significant fixed effect and a significant random effect was found. The fixed effect of −0.13 indicates that in general it is beneficial to plan early during the process. The correlation between task easiness and the random effect of the indicator was $r = -0.51$. The random effect shows that the effect of the delay indicator was near zero for difficult tasks and negative for easy tasks. This indicates that in easy tasks students who had their longest interval early during the process showed higher performance than students who had their longest interval later. For students working on difficult tasks this effect was less strong.

Table 3
Model parameters of Model1.2, Model2.2, and Model 3.2

Model1.2	Random effects:	Name	Variance	r		
Duration	Students:school	(Intercept)	0.47			
	School	(Intercept)	0.57			
	Task	(Intercept)	3.24			
		Duration	0.17	-.04		
	Fixed effects:	Estimate	SE	z-value	p-value	
	(Intercept)	0.22	0.40	0.54	0.589	
	Duration	−0.05	0.10	−0.47	0.639	
Model2.2	Random effects:	Name	Variance	r		
Delay	Students:school	(Intercept)	0.46			
	School	(Intercept)	0.56			
	Task	(Intercept)	3.35			
		Delay	0.02	-.51		
	Fixed effects:	Estimate	SE	z-value	p-value	
	(Intercept)	0.24	0.41	0.60	0.546	
	Delay	−0.13	0.05	−2.80	0.005	**
Model3.2	Random effects:	Name	Variance	r		
Variance	Students:school	(Intercept)	0.46			
	School	(Intercept)	0.56			
	Task	(Intercept)	3.42			
		Variance	16.78	-.39		
	Fixed effects:	Estimate	SE	z-value	p-value	
	(Intercept)	0.17	0.41	0.41	0.684	
	Variance	−1.85	1.31	−1.42	0.155	

Note. *p < 0.05, **p < 0.01, ***p < 0.001.

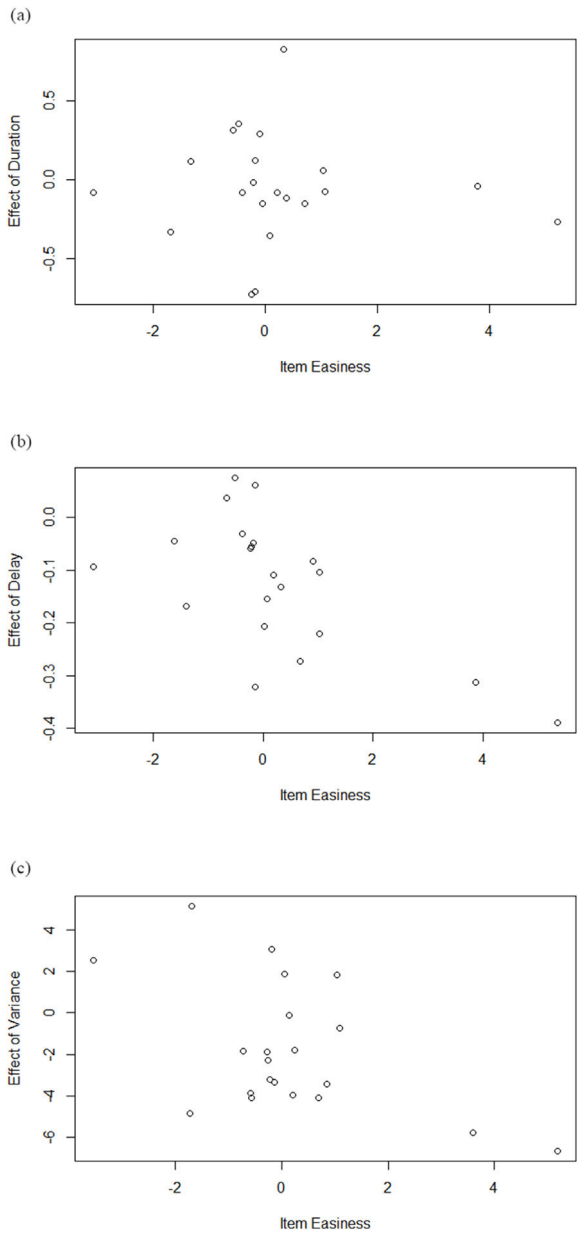

Fig. 2. Random effects of (a) the duration indicator, (b) the delay indicator, and (c) the variance indicator (calculated as fixed effect + task-specific effect) depending on task easiness (calculated as intercept + task-specific effect).

Research question 3: For the variance indicator no significant fixed effect but a significant random effect was found. The correlation between task easiness and the effect of the indicator was $r = -0.39$. The random effect indicates that in difficult tasks a high variation of time is beneficial while in easy tasks a low variation is beneficial.

Research question 4: The model including fixed effects of all three indicators and interactions between them is shown in Table 4. Significant effects were found for the interactions of duration-delay, delay-variance, and duration-delay-variance. Also the significant effect of the delay indicator was found again. The significant interaction effects are depicted in Fig. 3. The interaction between duration and delay shows that the duration indicator is positively related to success when the delay is high but has a very small negative effect when the delay is low (Fig. 3 (a)). The interaction between delay and variance shows that variance is positively related to success when the delay is low and negatively related to success when the delay is high (Fig. 3 (b)). The interaction effect of all three indicators shows that the effect of duration is not only depending on the delay but also on the combination of variance and delay (Fig. 3(c and d)). The positive effect of duration when the delay is high is largest when the variance is also high (Fig. 3 (c)). On the other hand, when the delay is low and the variance is high the effect of duration is negative (Fig. 3 (d)).

Table 4
Model parameters of Model 4.

Random effects:	Name	Variance	SD
Students:school	(Intercept)	0.46	0.68
School	(Intercept)	0.55	0.74
Task	(Intercept)	3.36	1.83

Fixed effects	Estimate	SE	*z*-value	*p*-value	
(Intercept)	0.14	0.41	0.34	0.735	
Duration	0.11	0.06	1.70	0.090	
Delay	−0.30	0.10	−3.12	0.002	**
Variance	−2.27	1.46	−1.56	0.120	
Duration*Delay	0.18	0.07	2.48	0.013	*
Duration*Variance	0.12	0.87	0.14	0.887	
Delay*Variance	−5.46	2.34	−2.33	0.020	*
Duration*Delay*Var.	2.73	1.30	2.09	0.036	*

Note. *p < 0.05, **p < 0.01, ***p < 0.001.

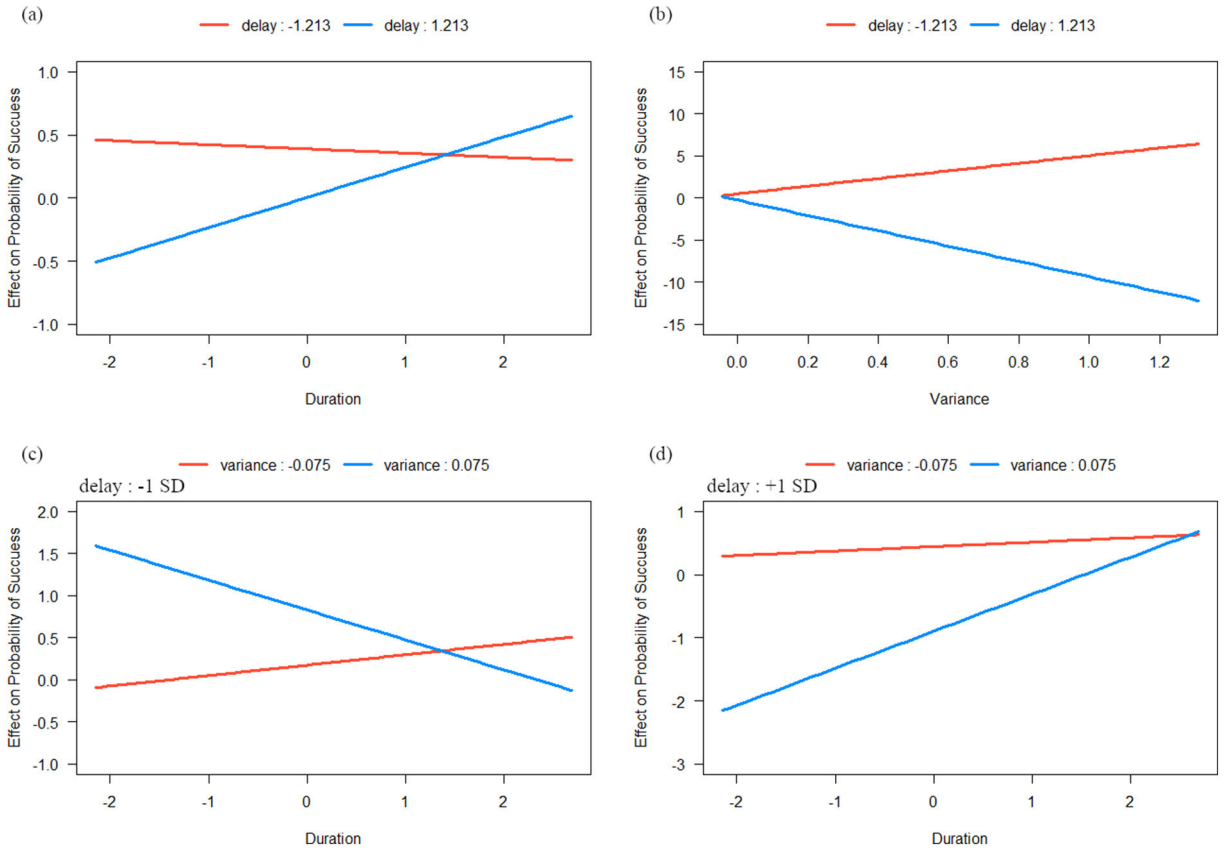

Fig. 3. Interaction effects between the centered variables (a) duration-delay, (b) delay-variance, (c) duration-variance with delay at +1SD, and (d) duration-variance with delay at -1SD.

4. Discussion

The aim of this study was to investigate the effect of planning behavior on performance in CPS. The results show that only one of our three planning indicators has a general effect on CPS performance, but that the effects of all three indicators differ between tasks. In other words, behavior that might be appropriate and highly beneficial in one task might actually be detrimental to performance in another task. Moreover, we found significant interaction effects of the three indicators.

In research question 1 we investigated whether the duration of planning intervals is related to performance. We used the duration of the longest time span between two successive interactions as an indicator for this behavior. The results show that the answer to this question depends on the task: in some tasks it seems to be disadvantageous to perform long planning intervals, while in other tasks

relatively long planning phases are beneficial. However, since the correlation between the random effects and task difficulty is rather low, task difficulty does not seem to be crucial for this effect. The PISA framework offers other task characteristics that might lead to such task specific effects such as amount of information, internal complexity, and reasoning skills required (OECD, 2013). For example, a task that has a high amount of information and a high internal complexity might require much planning and therefore benefit from a higher duration of the longest time span between interactions. On the other hand, in tasks that do not require much planning a high duration of the time span between interactions could - in line with the findings of Naumann et al. (2014) - indicate disorientation or mental overload.

In research question 2 we investigated whether the time when planning takes place is related to performance. For this question we used the delay of the longest interval between two interactions as a behavioral indicator. We found a significant negative effect of this indicator suggesting that in general planning should take place at an early stage of the process. Furthermore, the effect of this indicator is depending on task difficulty. In easy tasks it is beneficial to perform the longest planning phase early in the process, while in difficult tasks the effect of this indicator is very small. This supports the assumption that in easy tasks the whole process can and should be planned at an early stage of the task, while in difficult tasks planning ahead the entire process is not possible and, therefore, early planning is not beneficial. Hence, the mechanisms of planning in easy CPS tasks seem to be similar to those in non-complex problems. In these tasks a long initial planning phase was also beneficial (Albert & Steinberg, 2011; Unterrainer & Owen, 2006). In addition, the decrease of the effect in difficult tasks indicates that planning in these tasks can take place later in the process, since difficult tasks might require an initial phase of exploration before planning is useful. Therefore, initial planning seems to be particularly important in easier CPS tasks and non-complex problems.

In research question 3 we investigated whether the variation of planning time is related to performance in CPS. We wanted to know whether it is advantageous to concentrate all planning within a few phases or to spread it equally throughout the process. We used the variance of the times between the interactions as a behavioral indicator for this question. We did not find a general effect of this indicator but, again, we found a task-dependent effect that correlates with task difficulty. The effect is positive in difficult tasks and negative in easy tasks. Therefore, in easy tasks it seems to be important to distribute time equally throughout the process while in difficult tasks it is beneficial to engage in phases of higher and lower activity. This finding supports the assumption that planning is an important part of CPS that should appear at several stages of the process in difficult tasks, as stated in the framework of the OECD (2013).

In research question 4 we investigated whether there are any interaction effects between the three behavioral indicators. We found interaction effects of duration-delay, delay-variance, and duration-delay-variance showing a high interrelatedness of the effects of the three indicators. The interactions show that the effect of the duration of the longest interval depends on the delay: When the delay is high, the duration has a positive effect. When the delay is low the duration does not have an effect. This shows that the absence of early planning might be compensated by investing more time later in the process. The effect of variance also depends on the delay: When the delay is low, variance is positively related to success. When the delay is high, variance is negatively related to success. This indicates that an early planning phase that is longer than the later intervals is most beneficial. If students do not perform such an early planning phase they benefit from an equal time distribution among the whole process, so the absence of an early planning phase can partly be compensated by continued planning activities. Furthermore, the interaction between all three indicators shows that the effect of duration when delay is low also depends on variance. When students concentrate their time in an early planning phase this phase should be rather short. The reason for this negative effect of duration could be that students who take particularly long at the beginning of the task are students with low reading skills that therefore develop a poor understanding of the task. On the other hand, when performing at a steady speed the early planning time should be longer. In this case, students might benefit from an overall lower speed in task processing. On the other hand, when the longest planning period happens late and students work on a steady speed, the duration of planning does not play a role. However, when students concentrate their planning time in a late phase, this phase should be rather long. This could again compensate for the absence of planning at the beginning or during the prior process by investing time at a later stage. All in all, the optimal combination of planning behaviors would be a not too long planning interval at the beginning of the process that is followed by even shorter intervals.

5. Conclusion

To sum up our findings, we argue that the requirements for planning behavior in CPS are highly dependent on the task. The mechanisms in easy tasks seem to be similar to those of non-complex problems: a planning phase at the beginning of the process leads to higher performance. However, with increasing difficulty this effect changes. In difficult tasks, early planning has no benefit since these tasks might require exploration before meaningful planning is possible. In addition, in difficult tasks it is advisable to carefully allocate time at critical points, resulting in a high variation in the intervals between interactions. We also found interaction effects of our three indicators, so we assume that the effects of the three aspects of planning are highly intertwined. The optimal planning behavior identified through the interaction effects is a short planning interval at the beginning of the process that is followed by even shorter intervals.

On a more general level, our results show that the effects of different aspects of planning behavior in CPS tasks are very much interdependent. It seems that a generalization of beneficial behavior in CPS is difficult, since only one of the three behavioral indicators in our study had an overall effect on performance, but the effect of the other two indicators were depending on the values of the other indicators. Moreover, all indicators showed item-dependent effects. Therefore, the appropriateness of behaviors in CPS has to be regarded in the context of the particular task and in the context of other behaviors. One reason for the different requirements of tasks could be task characteristics such as those stated by the OECD (2013). All of these factors might affect the optimal

behavior in CPS.

However, this explanation does not stem from empirical findings and therefore the identification of task characteristics, which lead to the mentioned differences, should be subject to future research. Also, the operationalization of planning behavior we chose for our study is only one out of many possibilities, and of course has its weaknesses. As we only rely on time between interactions, it is not certain that students actually did spend this time planning. As we discussed earlier, the time between interactions is also related to other cognitive processes. Moreover, students could have performed other activities such as reading the task or engage in task-irrelevant behavior. Future research might address this issue using techniques like eye tracking or think aloud protocols to find out what mental states occur in students' minds while solving complex problems. Another limitation of this study lies in the method of reducing students' behavior to process indicators. Doing this a lot of information available in the log data is not included in the analysis and only a small proportion of the available information is used. A different approach to gain knowledge from log data is sequential pattern analysis (e.g. lag sequential analysis). Chang et al. (2017) used this method to identify advantageous behavioral patterns in a collaborative problem solving task. Since sequential pattern analysis allows identifying not only single types of behavior but behavioral patterns that are related to success it makes use of information about the whole problem solving process. However, the investigation of process indicators allows for inferences about the strength and shape of the relation between a specific behavior and success. Therefore, both methodological approaches seem suitable to identify beneficial behavior while solving problems.

As a consequence, process data gathered through computer-based assessment contributes to a better understanding of the process of CPS. We introduced three behavioral indicators, which reflect planning behavior in different ways and we showed that these indicators are related to students' performance in CPS tasks. We also showed that tasks require different strategies and that the effects of different behaviors depend on each other. In difficult tasks planning early in the process is less important than in easy tasks and the distribution of time into longer and shorter phases is advantageous. In easy tasks, early planning intervals and a steady process are beneficial. In general, having a short planning interval early in the process followed by more short intervals was identified as the most advantageous behavior. These findings can be used as a basis for both developing further behavioral indicators for cognitive processes and deepening our understanding of planning behavior in CPS.

Source of funding

This research was supported by German Federal Ministry of Education and Research [01LSA1504B].

Declarations of interest

None.

References

Albert, D., & Steinberg, L. (2011). Age differences in strategic planning as indexed by the tower of London. *Child Development, 82*(5), 1501–1517. https://doi.org/10.1111/j.1467-8624.2011.01613.x.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1)https://doi.org/10.18637/jss.v067.i01.

Baxter, G. P., & Glaser, R. (1997). An approach to analyzing the cognitive complexity of science performance assessments: CSE technical report 452. Los angeles. Retrieved from National Center for Research on Evaluation (Los Angeles), website https://www.cse.ucla.edu/products/reports/TECH452.pdf.

Blech, C., & Funke, J. (2005). Dynamis review: An overview about applications of the Dynamis approach in cognitive psychology. Retrieved from Deutsches Institut für Erwachsenenbildung website http://www.die-bonn.de/esprid/dokumente/doc-2005/blech05_01.pdf.

Blech, C., & Funke, J. (2010). You cannot have your cake and eat it, too: How induced goal conflicts affect complex problem solving. *The Open Psychology Journal, 3*, 42–53. Retrieved from http://cogprints.org/6867/1/Blech%26Funke_2010_Polytely.pdf.

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education, 53*(4), 1207–1217. https://doi.org/10.1016/j.compedu.2009.06.004.

Bransford, J., Brown, A. L., & Cockling, R. R. (1999). *How people learn: Brain, mind, experience, and school.* Washington, D.C: National Academy Press.

Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P. A. Frensch, & J. Funke (Eds.). *Complex problem solving: The European perspective* (pp. 27–63). Hillsdale, NJ: L. Erlbaum Associates.

Chang, C.-J., Chang, M.-H., Chiu, B.-C., Liu, C.-C., Fan Chiang, S.-H., Wen, C.-T., et al. (2017). An analysis of student collaborative problem solving activities mediated by collaborative simulations. *Computers & Education, 114*, 222–235. https://doi.org/10.1016/j.compedu.2017.07.008.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1–28. https://doi.org/10.18637/jss.v039.i12.

Dörner, D. (1989). *Die Logik des mißlingens: Strategisches Denken in komplexen situationen (17. - 19. Tsd).* Reinbek bei Hamburg: Rowohlt.

Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *The Journal of Problem Solving, 4*(1), 19–42. https://doi.org/10.7771/1932-6246.1118.

Frensch, P. A., & Funke, J. (1995). Definitions, traditions, and a general framework for understanding complex problem solving. In P. A. Frensch, & J. Funke (Eds.). *Complex problem solving: The European perspective* (pp. 24–43). Hillsdale, NJ: L. Erlbaum Associates.

Funke, J., & Frensch, P. A. (2007). Complex problem solving: The European perspective-10 Years after. In D. H. Jonassen (Ed.). *Learning to solve complex scientific problems* (pp. 25–47). New York: Lawrence Erlbaum.

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence, 3*(1), 21–40. https://doi.org/10.3390/jintelligence3010021.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626. https://doi.org/10.1037/a0034716.

Greiff, S., & Funke, J. (2008). *Indikatoren der Problemlöseleistung: Sinn und Unsinn verschiedener Berechnungsvorschriften: Bericht aus dem MicroDYN Projekt.*

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of

behavioral data from computer-generated log files. *Computers in Human Behavior, 61*, 36–46. https://doi.org/10.1016/j.chb.2016.02.095.
Klieme, E. (2004). Assessment of cross-curricular problem-solving competencies. In J. H. Moskowitz, M. Stephens, J. Moskowitz, & M. Stephens (Eds.). *Comparing learning outcomes: International assessment and education policy* (pp. 81–107). London: Routledge Falmer.
Lesh, R., & Zawojewski, J. (2007). Problem solving and modeling. In F. K. Lester (Ed.). *The handbook of research on mathematics teaching and learning* (pp. 763–804). .
Leutner, D., Funke, J., Klieme, E., & Wirth, J. (2005). Problemlösefähigkeit als fächerübergreifende Kompetenz. In E. Klieme, D. Leutner, & J. Wirth (Eds.). *Problemlösekompetenz von Schülerinnen und Schülern: Diagnostische Ansätze, theoretische Grundlagen und empirische Befunde der deutschen PISA-2000-Studie* (pp. 11–19). (1st ed.). Wiesbaden: VS Verl. für Sozialwiss.
Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander, & P. H. Winne (Eds.). *Handbook of educational psychology* (pp. 287–304). (2nd ed.). Mahwah, N.J: Erlbaum.
Naumann, J., & Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences, 53*, 1–16. https://doi.org/10.1016/j.lindif.2016.10.002.
Naumann, J., Goldhammer, F., Rölke, H., & Stelter, A. (2014). Erfolgreiches Problemlösen in technologiebasierten Umgebungen: Wechselwirkungen zwischen Interaktionsschritten und Aufgabenanforderungen. *Zeitschrift für Padagogische Psychologie, 28*(4), 193–203. https://doi.org/10.1024/1010-0652/a000134.
Novick, L. R., & Bassok, M. (2005). Problem solving. In K. J. Holyoak, & R. G. Morrison (Eds.). *The cambridge handbook of thinking and reasoning* (pp. 321–349). Cambridge: Cambridge University Press.
OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* PISAParis: OECD.
OECD (2015). *Students, computers and learning: Making the connection.* Paris: OECD.
Osman, M. (2010). Controlling uncertainty: A review of human behavior in complex dynamic environments. *Psychological Bulletin, 136*(1), 65–86. https://doi.org/10.1037/a0017815.
R Core Team (2016). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.
Reeff, J., Zabal, A., & Blech, C. (2006). The assessment of problem-solving competencies: A draft version of a general framework. Retrieved from Deutsches Institut für Erwachsenenbildung website http://www.die-bonn.de/esprid/dokumente/doc-2006/reeff06_01.pdf.
Rouet, J.-F., & Le Bigot, L. (2007). Effects of academic training on metatextual knowledge and hypertext navigation. *Metacognition and Learning, 2*(2–3), 157–168. https://doi.org/10.1007/s11409-007-9011-z.
Unterrainer, J. M., & Owen, A. M. (2006). Planning and problem solving: From neuropsychology to functional neuroimaging. *Journal of Physiology Paris, 99*(4–6), 308–317. https://doi.org/10.1016/j.jphysparis.2006.03.014.
Vosniadou, S., & Ortony, A. (1989). *Similarity and analogical reasoning.* Cambridge, New York: Cambridge University Press.
Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy & Practice, 10*(3), 329–345. https://doi.org/10.1080/0969594032000148172.

ANHANG B: ARBEIT 2

Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2020). Using Process Data to Explain Group Differences in Complex Problem Solving. *Journal of Educational Psychology*. Advance online publication. http://dx.doi.org/10.1037/edu0000446

Using Process Data to Explain Group Differences in Complex Problem Solving

Beate Eichmann[a], Frank Goldhammer[a], Samuel Greiff[b], Liene Brandhuber[c], Johannes Naumann[d]

a: DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student Assessment (ZIB)

b: University of Luxembourg

c: Goethe University Frankfurt

d: University of Wuppertal

Author Note

Corresponding author: Beate Eichmann, DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student Assessment (ZIB) Rostocker Straße 6, 60323 Frankfurt am Main, Germany, Tel: +49 69 24708-863, beate.eichmann@dipf.de

This research was funded by the German Federal Ministry of Education and Research (grant numbers: 01LSA1504A and 01LSA1504B) and by a project funded by the Fonds National de la Recherche Luxembourg (The Training of Complex Problem Solving; "TRIOPS").

Parts of this research were presented at the NCME 2018 Annual Meeting and at the GEBF 2018 Annual Meeting.

Submitted: March 19, 2019

Abstract

In large scale assessments, performance differences across different groups are regularly found. These group differences (e.g. gender differences) are often relevant for educational policy decisions and measures. However, the formation of these group differences usually remains unclear. We propose an approach for investigating this formation by considering behavioral process measures as mediating variables between group membership and performance on the 2012 PISA complex problem solving items. We found that across all investigated countries interactive behavior can fully explain gender differences in CPS, but cannot explain differences between students with and without a migration background. However, in some countries these results differ from the cross-country results. Our results indicate that process measures derived from log data are useful for further investigating and explaining performance differences between girls and boys and students with and without migration background.

Using Process Data to Explain Group Differences in Complex Problem Solving

In educational settings, performance differences between demographic groups are regularly observed (OECD, 2011, OECD, 2014a; Prenzel et al., 2004). Group variables that are associated with performance differences include students' gender or migration background. For instance, in large scale assessments like the Programme for International Student Assessment (PISA), girls regularly outperform boys in the domain of reading, and boys regularly outperform girls in mathematics (OECD, 2011, OECD, 2014a). These performance differences raise questions of equity and fairness in the educational systems in which they occur, which sometimes leads to changes in educational policy. For instance, after the results of PISA 2000 were published, national educational standards were established in Germany to reduce performance differences associated with social background (Neumann, Fischer, & Kauertz, 2010). To understand the causes of and eventually minimize these performance differences, it is important to investigate their underlying mechanisms. For instance, in the domain of literacy, Artelt, Naumann, and Schneider (2010) found that girls outperform boys in reading partly due to a better command of metacognitive strategies.

Another domain where performance differences between demographic groups arise regularly is problem solving. For example, in PISA 2012, several group-level performance differences in problem solving were observed across OECD countries (OECD, 2014b). One factor related to problem solving performance was gender; boys scored significantly higher than girls. The difference was 0.07 standard deviations on average across countries. Another factor related to problem solving ability was migration background; students without a migration background scored significantly higher than students with a migration background. On average, this difference amounted to 0.32 standard deviations (OECD, 2014b).

Complex problem solving

A problem is defined as a situation in which a person wants to achieve a goal but no obvious solution is available (Mayer & Wittrock, 2006). Problems can further be divided into the subdomains of analytical (static) and complex (dynamic) problems. Complex problem solving (CPS) is a particularly important competency in today's society and can be regarded as a 21st century skill since it is required in many situations in everyday life (Binkley et al., 2012). In contrast to analytical problems, complex problems change dynamically (over time or as a result of interaction), and not all information required to solve the problem is present at the outset (Frensch & Funke, 1995). Therefore, complex problems require the problem solver to interact with the problem to obtain necessary information, while analytical problems can be solved just by thinking. Due to its interactive nature, CPS is usually measured via computer-based assessment. In computer-based assessment, log data is available that can give insight into the task solution process (Greiff, Wüstenberg, & Avvisati, 2015; Kroehne & Goldhammer, 2018). This means that differences between groups in terms of cognitive and behavioral task engagement can be analyzed as one possible source of differences in task success, and thus in estimated ability. In the present study, we investigate whether performance differences in CPS between demographic groups can be explained by behavioral differences extracted from log data.

Problem solving process: Interaction and exploration behavior

What behavioral processes might account for individual success, and thus possibly also group differences, in CPS? Some predictions can be made on the basis of the literature on error management training. Error management training refers to an active learning approach in which trainees are encouraged to explore a system, even if this means they might commit errors. The core elements of error management training are experimentation and exploration without

providing much guidance. Trainees are supposed to commit errors to learn how to handle the system and errors within the system (see Frese & Keith, 2015 for an overview). Error management training is often used to train people to use software systems. A number of studies have shown that error management training is superior to error-avoidant training, especially in novel situations that require knowledge transfer, and thus in situations which could also be characterized as problems (e.g. Bell & Kozlowski, 2008; Keith & Frese, 2005; Keith, Richter, & Naumann, 2010; see Keith & Frese, 2008, for a meta-analysis). Dormann and Frese (1994) investigated the effect of exploration in error management training on performance in a complex technology-based environment. Participants received either error-avoidant training or error training within a statistics program before their performance was assessed. While the subjects in the error-avoidant group received a structured tutorial that prescribed every step necessary to accomplish specific tasks with the statistics program, the error training group only received a set of basic commands and no further introduction. Participants were selected to have basic statistical experience and basic experience with similar software. The study's authors defined exploration during training as interactions with the environment that had not been previously introduced. They found higher performance in the error training group than in the error-avoidant group and a positive effect of exploration in the training session on subsequent performance. The superior performance of the error training group was shown to be attributable to trainees in the error training engaging in more comprehensive exploration, and thereby acquiring a deeper understanding and becoming familiar with more possible states of the system (Dormann & Frese, 1994). However, Dormann and Frese (1994) state that exploration should not be confused with trial-and-error, since exploration should be guided by hypotheses based on a mental model of the system. Kapur (2008) also argues that exploration leads to higher performance in problem

solving. He showed that exploration led to a better representation of and to higher knowledge about the problem. Bell and Kozlowski (2008) also reported a positive effect of exploration in training on subsequent performance in a complex computer-based simulation. They compared the performance of university students instructed to use exploration in a training phase with those who received a structured step-by-step tutorial. Students in the exploration group were instructed to explore the system to discover suitable methods and strategies for handling it, while students in the other group received detailed step-by-step instructions on how to accomplish goals and were asked to follow these instructions. Both groups were provided with the same list of training objectives to achieve. The students in the exploration group exhibited higher performance in the follow-up assessment than students who received the structured tutorial. Bell and Kozlowski (2008) argue that exploration provides learners with control over their learning process, which in turn activates their metacognition (e.g. planning, monitoring and revising behavior). As both Bell and Kozlowski (2008), and Keith and Frese (2005) argue, metacognitive processes activated by exploration enhance learning and transfer. For example, Bell and Kozlowski (2008) argue that self-evaluation activities are positively related to participants' strategic knowledge, intrinsic motivation and self-efficacy. Drawing upon these perspectives, we argue that explorative behavior is a crucial prerequisite for successful CPS, since the cognitive processes mentioned above (e.g. activation of metacognition, problem representation, hypotheses testing) also apply to CPS.

Besides exploration, the overall amount of interaction with a problem scenario might also predict CPS performance. Interaction with a problem means that the problem solver takes any observable action. In a computer-based scenario this could be, for example, a mouse click or a keystroke. Naumann, Goldhammer, Rölke, and Stelter (2014) investigated effects of interaction

on problem solving success in technology-based environments using data from the Programme for the International Assessment of Adult Competencies (PIAAC) (see OECD., 2013b). They found that the number of interactions had a quadratic relation with problem solving performance. The inverse u-shape had its optimum at 1.5 standard deviations above average; thus, it seems that students who refrain from interaction particularly struggle. Naumann et al. (2014) argue that low computer-related self-efficacy or high computer-related anxiety might be the reasons for persons to behave passively and therefore be less successful in solving technology-based problems. Since in PIAAC the participants' basic computer skills were assessed and the computer-based problem solving test was only administered to those participants who showed sufficient computer skills, the unfamiliarity with computers is unlikely to be the reason for the passive behavior of some participants. On the basis of their results, Naumann et al. (2014) argue that in problem-solving in technology-based environments, acting too cautiously might pose a greater threat to performance than acting too boldly. This interpretation is in line with the literature on error management training, where training approaches that encourage "risky" behavior when learning – i.e. willingness to commit an error, and learn from it – is recommended (Bell & Kozlowski, 2008; Dormann & Frese, 1994).

In the present study, we distinguish between interaction behavior and exploration behavior, with exploration representing a subset of overall interactions. More specifically, we define interaction as a student's engagement with a problem irrespective of what exactly they do. Further, we define exploration as a student's interaction with aspects of a problem that provide information about the problem situation but do not directly contribute to its solution. For instance, if the problem requires the student to buy a daily ticket for the subway, clicking on the button for individual trips on the ticket machine does not provide an instant solution and

therefore would be defined as exploration behavior, as it might help the student build a better representation of the problem space. Repeated interaction with solution-relevant aspects of a problem is also regarded as exploration as long as it is not essential for solving the problem. Hence, if the student in the example above would go back and press the button for a daily ticket again after having already pressed it, the second button press would be regarded as exploration (because the student could have already solved the problem after the first button press). Conversely, the first press on the button for a daily ticket is classified as an instance of interaction, but not exploration (since this button press is directly goal-oriented).

Explaining group differences in CPS

Both theory and empirical results point to the importance of the amount of interaction and exploration in CPS on the individual level (Bell & Kozlowski, 2008; Dormann & Frese, 1994; Frensch & Funke, 1995; Naumann et al., 2014). Building upon this perspective, in the present research, we ask whether the amount of interaction and exploration can also account for CPS performance differences *between groups* where they exist. Note that whether group-level differences in *complex* problem solving exist at all is not a trivial question. PISA 2012 demonstrated differences between boys and girls, and between students with and without a migration background in overall problem solving, that is, in a measure of problem solving that entailed both complex *and* analytical problem solving. On this basis, it is plausible to assume that similar performance differences would emerge if complex problem solving would be specifically examined.

If indeed performance differences in CPS exist between girls and boys, these might be traced back to behavioral differences associated with gender. Wittmann and Hattrup (2004) found behavioral differences between females and males that might account for gender differences in

CPS performance. They argue that males outperformed females because males engage in more risky behavior, which in turn provoked more dramatic changes in the CPS scenario, and in consequence, provided them with more information and more learning opportunities about the system. These findings are in line with the results of a meta-analysis by Cross, Copping, and Campbell (2011) indicating that males in general exhibited more risk-taking behavior than females. Wüstenberg, Greiff, Molnár, and Funke (2014) also found higher CPS performance among male students compared to females. They observed that boys applied the VOTAT strategy (VOTAT=vary one thing at a time) associated with high CPS performance more often. Wüstenberg et al. (2014) found similar differences in the application of the strategy between girls and boys in different countries. Sonnleitner, Brunner, Keller, and Martin (2014) investigated behavioral differences in CPS between students with and without a migration background. They found that students with a migration background exhibited more exploration behavior than their peers without a migration background. However, students without a migration background exhibited higher CPS performance. Sonnleitner et al. (2014) explained this result by arguing that students with a migration background might have had difficulties transferring the generated information into declarative knowledge. Martin, Liem, Mok, and Xu (2012) identified socio-economic status, language background, age of migration, gender, and age as factors that are relevant to immigrant students' problem solving performance. From the available evidence, there seems to be good grounds to assume that differences in interaction and exploration might indeed account for performance differences in CPS between boys and girls. In contrast, there is a much weaker basis for this assumption regarding performance differences in CPS between students with and without migration.

Thus, the first aim of the present study was to investigate whether the performance differences between boys and girls and between students with and without a migration background that PISA 2012 found for overall problem solving would also be present in the CPS subdomain. If this would be the case, the second aim was to investigate whether these effects are mediated through behavioral patterns while solving complex problems (OECD, 2014b), specifically interaction and exploration.

Hypotheses

Performance differences between groups

On the basis of previous results demonstrating performance differences in CPS between boys and girls (Wüstenberg et al., 2014) and students with and without a migration background (Sonnleitner et al., 2014), we expect to find performance differences between these groups as well:

Hypothesis 1a: Boys exhibit higher CPS performance than girls.

Hypothesis 1b: Students without a migration background exhibit higher CPS performance than students with a migration background.

Effects of behavioral indicators on CPS performance

As Frensch and Funke (1995) claim, CPS requires interacting with the problem. Moreover, the amount of interaction can predict performance in technology-based problem solving (Naumann et al., 2014). The work of Bell and Kozlowski (2008) and Dormann and Frese (1994) showed that exploration behavior has a positive impact on performance in complex environments that have a problem-like character. Therefore, we hypothesize that the amount of interaction and exploration is also related to CPS performance:

Hypothesis 2a: The number of interactions is positively related to CPS performance.

Hypothesis 2b: The number of exploration steps is positively related to CPS performance.

Behavioral differences between groups

As mentioned above, Wittmann and Hattrup (2004) and Wüstenberg et al. (2014) found behavioral differences between females and males when solving complex problems. The former observed more risky behavior in males, while the latter observed that males apply the VOTAT strategy more often. In both studies, boys exhibited higher performance and more behavior associated with high performance. Both risky behavior and the VOTAT strategy might be associated with exploration. When exploring, students risk committing errors. The VOTAT strategy requires students to explore a problem scenario to obtain information required for solving the problem. Therefore, we expect boys to engage in more exploration than girls, and since exploration is part of interaction, we also expect boys to engage in more interaction than girls.

The findings by Sonnleitner et al. (2014) indicate that students with a migration background engage in more exploration behavior than their fellow students without a migration background. Therefore, we assume that students with a migration background engage in more exploration and thus also (since exploration is part of interaction) more interaction than students without a migration background.

Hypothesis 3a: Boys exhibit more interactive and explorative behavior than girls.

Hypothesis 3b: Students with a migration background exhibit more interactive and explorative behavior than students without a migration background.

Mediation of performance differences

From previous theory and findings, there is ample reason to hypothesize that performance differences in complex problem solving between boys and girls might be mediated through

different styles of engagement, and specifically through interaction and exploration. Boys tend to be more prone to engage in "risky" behavior (Wittmann & Hattrup, 2004), and thus to engage in exploration, which is a useful strategy in complex problem solving:

Hypothesis 4a: The effect of gender on complex problem solving performance is mediated through exploration behavior.

Since, in addition to exploration, previous research has shown that interaction in the sense of both directly task-related behavior as well as exploratory behavior is mostly positively related to CPS performance (Naumann et al., 2014), interaction might also be suspected to mediate the effects of gender:

Hypothesis 4b: The effect of gender on complex problem solving performance is mediated through interaction behavior.

In contrast to gender, the role of behavioral differences between students with and without a migration background in bringing about performance differences between the two groups is much less clear. For example, while students with a migration background lag behind their peers without a migration background, they tend to do *more* exploration (Sonnleitner et al., 2014), behavior that is considered here to be beneficial for complex problem solving. Thus, while differences in the behavioral indicators considered here might account for CPS performance differences related to migration background, other variables like language proficiency might be even more crucial (Martin et al., 2012). For these reasons, we refrain from specifying a hypothesis regarding the mediating role of exploration and interaction with respect to the effects of migration background. Instead, we treat the question of whether the effects of migration background on CPS performance are mediated through exploration and interaction as an exploratory research question.

Method

Sample

We used log data generated during the computer-based assessment of problem solving in PISA 2012. We only used data from the 44 countries that participated in the computer-based assessment of problem solving. However, of those countries, we had to exclude Cyprus because data on gender and migration status were not available. We also decided to exclude Korea since not all of our models converged with the Korean dataset. Therefore, the overall sample size was further reduced to N=81,039 students from 42 countries (50.15% female, 12.22% with a migration background). The number of participating students, the percentage of female students and the percentage of students with a migration background in each country are shown in Appendix A.

Instruments

The CPS assessment in PISA 2012 consisted of 27 computer-based items, which were organized into 16 units alongside analytical problem solving items. The conceptual distinction between complex (called interactive in PISA) and analytical items (called static in PISA) was made by the OECD (2013a). Their criterion to distinguish between the two item types was the disclosure of information about the problem. In contrast to static problems, in interactive problems not all information was disclosed at the outset to the problem solver. Therefore, we argue that this definition of interactivity matches the definition of CPS by Frensch and Funke (1995) we follow in this article. Each unit was comprised of two to three items with similar stimulus material. Students worked on one or two out of four different problem solving clusters, which included four units each. The item order within units was always the same. After finishing an item, students could not return to it.

The items were embedded in everyday contexts, and they were designed to control for prior knowledge by applying very heterogeneous contexts (OECD, 2013a). Moreover, prior knowledge was not sufficient to solve a problem solving item in PISA 2012. Examples of these everyday contexts are controlling room temperature and humidity using a climate control panel and buying train tickets from a ticket machine. Prior knowledge about ticket machines might help to solve this item, however the item still requires CPS activities to be solved. In principal, prior knowledge could have been avoided by constructing abstract or artificial contexts for a task. However, problem solving items without a meaningful context might lack external validity to everyday problems and, therefore, lack relevance. Moreover, in PISA 2012 Differential Item Functioning analyses were carried out to make sure that all items worked equally well for students of different gender, of different language proficiency or from different countries. In this procedure problematic items were revised or excluded (OECD, 2014c).

The computer-based CPS items required only basic computer skills like clicking on virtual buttons and sliders, dragging and dropping, operating simulated machines, exploring simulated environments, and manipulating variables. The response formats included simple and complex multiple-choice items that were answered by clicking radio buttons (26% of the items), items in which selections had to be made from pull-down menus (7% of the items), items that required parts of diagrams to be drawn (26% of the items), items that required establishing a certain state by clicking buttons (37% of the items) and text boxes (11% of the items) (OECD, 2013a). Note that some items contained more than one response format. Before the assessment, a tutorial was administered so that students could practice the required skills. The tutorial was offered to all students. However, students could choose whether they wanted to work on the tutorial or skip it. For a detailed explanation of the item characteristics, see OECD (2013a).

Gender and migration status were assessed via a student questionnaire. Students were asked whether they or their parents were born abroad. As recommended by OECD (2014c), we defined a migration background as having two parents who were born outside the country of assessment. We did not differentiate between first and second generation migration background.

Scoring

We used the available log data to score students' responses. For CPS performance, the response coding suggested by OECD (2015b) was used: The responses were coded as either correct (1) or incorrect (0). Therefore, our CPS scoring indicates whether or not a problem was completely solved. The following example should illustrate the scoring: The problem to be solved is to buy the cheapest ticket from a ticket machine that would meet certain criteria (e.g. a ticket for several trips with the city subway including a student discount). Since the problem solvers are not familiar with the ticket machine they cannot know what fares and ticket types are available and how to choose between them. This item was coded as being correctly solved if the problem solver bought the ticket that met all the criteria specified in the task. If the problem solver bought a different ticket or did not buy a ticket at all, the item would be coded as not correctly solved. To rule out any confounds between our process measures and CPS performance, we did not use the original PISA scoring rules that included scores for students' behavior during the task. Out of the 27 CPS items, two items were excluded because the respective log files included clicks on elements with ambiguous IDs. Another two items had to be excluded because the responses included free text input, which was not recorded in the log data, so the correctness of the response could not be inferred. Four items were excluded because the total number of user interactions (which is relevant for our analyses) was externally restricted. Another three items were excluded because interaction with the item was

automatically stopped as soon as the correct system state was reached, also limiting the possibilities for exploration. After these exclusions, a total of 16 items remained out of the 27 CPS items (for item examples see OECD, 2013a).

We extracted two process measures from the log data: (1) the overall number of interactions and (2) the number of exploration steps. For the overall number of interactions, we counted all click events that occurred for each item completion sequence. To determine the number of exploration steps, we defined the shortest possible click pattern that would lead to a correct solution for each item. For some items, more than one shortest click pattern was defined, since there were multiple equally short ways to solve the item. We then calculated the Levenshtein distance (LD) between these shortest click patterns and the students' actual click patterns. The LD is a measure of the difference between two sequences. It counts the number of insertions, deletions and substitutions of sequence elements necessary to transform one sequence into the other (Navarro, 2001). For items with more than one shortest click pattern, the LDs between every shortest pattern and the students' patterns were calculated, meaning that every student had several LDs for these items. Then, we took the minimum LD for each student for further analysis to obtain a conservative estimate of the amount of exploration. When a student's click pattern is identical to the shortest possible pattern, the LD equals zero. Hence, an LD of zero indicates that no exploration took place. The LD increases with an increasing number of clicks. However, the LD also increases with a decreasing number of clicks for click patterns shorter than the shortest possible solution pattern. To get a valid measure for the number of exploration steps, we therefore adjusted the LD for those shorter click patterns by subtracting the difference between the length of the shortest possible pattern and the length of the actual click pattern. This gave us an indicator of the number of exploration steps that counted every

interaction beyond the "minimal" item solution, i.e. an item solution process that does not entail any kind of exploratory behavior. We expect the distribution of the number of exploration steps to be right-skewed, since it is a mere count of students' exploration steps. Table 1 provides an example of how the number of exploration steps was computed. In the example the task has two possible shortest click patterns to come to the correct solution (pattern 1: A-B-C, pattern 2: A-D-E-F). Student 1 has a click pattern longer than these shortest patterns (A-D-E-F-G-H). Therefore, no adjustment is necessary. Student 1's number of exploration steps equals the minimum LD between his pattern and the two shortest patterns (which is two). However, student 2 hast a click pattern shorter than the shortest possible patterns (A-D). Therefore, the LD needs to be adjusted. The adjustment is accomplished by subtracting the difference between the length of student 2's pattern and the respective shortest patterns (i.e. -1 for pattern 1 and -2 for pattern 2). After the adjustment, the minimum LD equals the number of student 2's exploration steps (which is zero).

Table 1: Example calculation of the number of exploratory steps for an item with two possible shortest click patterns for two different students. LD is only adjusted for student 2, since this student's click pattern is shorter than the shortest possible pattern.

	Click pattern student 1: A-D-E-F-G-H	Click pattern student 2: A-D
Levenshtein dictance without adjustment:		
Shortest pattern 1: A-B-C	5	2
Shortest pattern 2: A-D-E-F	2	2
Levenshtein dictance after adjustment:		
Shortest pattern 1: A-B-C	5	1
Shortest pattern 2: A-D-E-F	2	0
Number of exploration steps	2	0

Procedure

PISA 2012 covered mathematics, reading, science, problem solving, and financial literacy. Problem solving, mathematics and reading were assessed via computer-based assessment. Students were administered two computer-based clusters of which none, one, or both

were problem solving clusters. Since the problem solving clusters were randomly assigned to the students, parameters should be comparable across items. Also the pairs of clusters overlapped meaning that all possible combinations of item clusters were administered. The computer-based assessment took place after the completion of the paper-based PISA items. Students had 20 minutes to complete each cluster (OECD, 2014b).

Data Preparation

Data preparation was performed separately for every country. First, the two process indicators were inspected for outliers. An outlier was defined as a value three standard deviations above/below the respective average. The outliers were replaced with the value exactly three standard deviations above/below the average, as suggested by Goldhammer et al. (2014). We identified outliers corresponding to 1.08% of the values for the number of interactions and 2.71% of the values for the amount of exploration. Because both indicators showed high skewness, we log-transformed the indicators after adding +1 to every value to enable logarithmic transformation of zeros. The log transformation was necessary to normalize the data in order to perform a maximum likelihood estimation. By log transforming we changed the metric of our data leading to a numeric reduction of differences in the upper value range.

Data Analysis

We estimated four mediation models for each country. In each model, the independent variable was either gender or migration background, and the mediating variable was either interaction or exploration. The criterion variable in each case was CPS performance, defined as correctly solving the respective problems. Both gender and migration background were dichotomous (male=0, female=1; no migration background=0, migration background=1). The mediator variables were modeled as latent variables on the person level, aggregating the

respective process indicator across all CPS items. The item intercepts were estimated freely to account for different demands for interactions across items. The dependent variable (CPS performance) was also modeled as a latent variable on the person level, aggregating the scores across all CPS items. . In all our models we controlled for reading ability as a possible confound. Since reading ability is required to understand and answer PISA's CPS tasks properly, students with migration background might experience higher difficulty due to lower language ability. Therefore, Martin et al. (2012) suggest controlling for reading ability when investigating CPS among students with migration background. But also between girls and boys differences in reading ability are regularly observed which might decrease boys' performance in CPS tasks (OECD, 2014a). Therefore, we decided to control for reading ability in all our analyses. We used weighted likelihood estimators based on the items from the PISA 2012 paper-based reading assessment to control for reading ability in the mediator variables and in CPS performance. The resulting models are shown in Figure 1. To account for PISA's complex sampling design, the final PISA student weights were used and school was included as a cluster variable. We used Mplus 7.4 for the mediation analyses (Muthén & Muthén, 1998-2015) and the R package MplusAutomation for executing the country-specific calculations (Hallquist & Wiley, 2018; R Core Team, 2016). The effect size κ^2 was calculated for the mediation effect of every country using the R package MBESS (Kelley & Lai, 2010). Following procedures suggested by Naumann (2015, 2019), we employed random-effects meta-analysis (see Hedges & Vevea, 1998) to integrate country-wise results, providing both an estimate of a fixed effect that is the same across countries, as well as an estimate of each effects variance across countries. We used the R package metafor (Viechtbauer, 2010) to summarize the mediation model parameters across countries, treating each country as a "study" in a meta-analysis. As a measure of between

country variation, we report the square root of estimated between-country variance τ, the Q-statistic for testing τ^2 for heterogeneity, and its p-value. The effect size κ^2 was aggregated using its median.

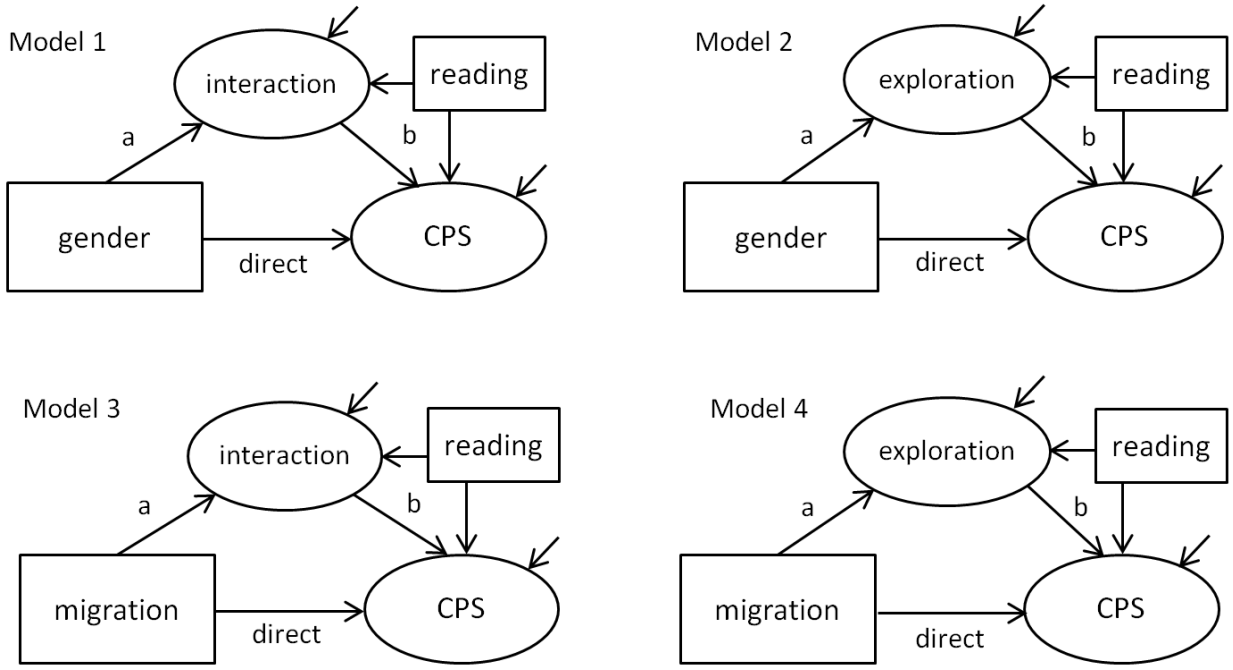

Figure 1: The four resulting models.

Results

Table 2 shows the aggregated standardized model results and effect sizes. The model results and effect sizes by country are listed in Appendix B, Appendix C, Appendix D, and Appendix E. An overview of the estimated models is given in Figure 1.

Performance differences between groups (Hypothesis 1a and 1b)

We found a significant total effect of gender on CPS performance, indicating that boys outperformed girls in both our models incorporating gender (Model 1: total effect=-0.28, $p<.001$; Model 2: total effect=-0.28, $p<.001$). This supports Hypothesis 1a. Also as expected, there was a total effect of migration background on CPS performance indicating that students without a migration background outperformed students with a migration background (Model 3: total effect

=-0.16, p=.001; Model 4: total effect =-0.12, p=.003). Therefore, Hypothesis 1b was supported as well. However, both observed performance differences varied significantly between countries (Model 1: τ(total effect)=0.09, p<.001; Model 2: τ(total effect)=0.09, p=.002; Model 3: τ(total effect)=0.22, p<.001; Model 4: τ(total effect)=0.19, p<.001). Although the aforementioned performance differences were found in most countries, in 1 country the effect of gender and in 11 countries the effect of migration background on performance was reversed. For example, in the United Arab Emirates girls outperformed boys and in Poland immigrant students outperformed students without a migration background (see Appendix B, Appendix C, Appendix D, Appendix E).

Table 2: Aggregated model estimates and effect sizes

	parameter	estimate	SE	z	p	τ	Q(41)	p
Model 1	a	-0.27	0.02	-12.32	<.001	0.12	135.18	<.001
	b	0.71	0.01	62.86	<.001	0.06	100.18	<.001
	total	-0.28	0.02	-14.24	<.001	0.09	99.40	<.001
	direct	-0.08	0.01	-6.42	<.001	0.04	46.99	.241
	indirect	-0.19	0.01	-13.25	<.001	0.07	98.26	<.001
	κ^2	0.13						
Model 2	a	-0.57	0.02	-27.14	<.001	0.09	71.88	.002
	b	0.44	0.02	30.32	<.001	0.03	43.76	.355
	total	-0.28	0.02	-14.44	<.001	0.09	98.84	.002
	direct	-0.03	0.02	-1.73	.083	0.05	45.44	.292
	indirect	-0.23	0.01	-19.68	<.001	0.03	43.08	.382
	κ^2	0.17						
Model 3	a	-0.09	0.05	-1.67	.094	0.27	182.00	<.001
	b	0.74	0.01	60.95	<.001	0.06	106.92	<.001
	total	-0.16	0.05	-3.45	.001	0.22	154.70	<.001
	direct	-0.07	0.05	-1.45	.146	0.27	199.50	<.001
	indirect	-0.07	0.04	-1.77	.077	0.19	172.56	<.001
	κ^2	0.05						
Model 4	a	0.02	0.05	0.43	.665	0.25	141.55	<.001
	b	0.44	0.02	29.76	<.001	0.05	56.72	.052
	total	-0.12	0.04	-2.97	.003	0.19	122.96	<.001
	direct	-0.15	0.04	-4.33	<.001	0.15	79.51	<.001
	indirect	0.02	0.02	1.19	.235	0.08	87.96	<.001
	κ^2	0.03						

Note. Model 1 and 2: independent variable is gender. Model 3 and 4: independent variable is migration background. Model 1 and 3: mediator is number of interactions. Model 2 and 4: mediator is amount of exploration. Parameter a represents the effect of the independent variable on the mediator. Parameter b represents the effect of the mediator on the dependent variable.

Effects of behavior indicators on CPS performance (Hypothesis 2a and 2b)

The results from Model 1 and Model 3 showed that the number of interactions was positively related to CPS performance (Model 1: b=0.71, p<.001; Model 3: b=0.74, p<.001). The number of exploration steps was also positively related to CPS performance (Model 2: b=0.44, p<.001; Model 4: b=0.44, p<.001). Thus, Hypothesis 2a and Hypothesis 2b were supported. The effect of the number of interactions exhibited significant between-country variation, while the effect of exploration did not (Model 1: τ(b)=0.06, p<.001; Model 3: τ(b)=0.06, p<.001; Model 2: τ(b)=0.03, p=.355; Model 4: τ(b)=0.05, p=.052). However, in all countries both effects pointed in the same direction (see Appendix B, Appendix C, Appendix D, Appendix E).

Behavioral differences between groups (Hypothesis 3a and 3b)

Gender. The results of Model 1 and Model 2 showed that gender was negatively related to the number of interactions and the number of exploration steps, indicating that boys exhibited more interactions and more exploration than girls (Model 1: a=-0.27, p<.001; Model 2: a=-0.57, p<.001), providing support to Hypothesis 3a. The relation between gender and exploration was stronger than the relation between gender and interactions. We found significant between-country variations for both effects (Model 1: τ(a)=0.12, p<.001; Model 2: τ(a)=0.09, p=.002). The estimates pointed in the same direction in all countries for the number of exploration steps and in all countries except one for the number of interactions (see Appendix B, Appendix C). However, in the United Arab Emirates the relation between gender and number of interactions was reversed.

Migration Background. We did not find a significant association between migration background and number of interactions (Model 3: a=-0.09, p=.094). However, we found significant between-country variation (Model 3: τ(a)=0.27, p<.001). We did not find a relation

between migration background and the amount of exploration (Model 4: a=0.02, p=.665). Once again, we found significant between-country variation of this effect (Model 4: τ(a)=0.25, p<.001). Therefore, Hypothesis 3b is not supported.

Mediation of performance differences (Hypothesis 4a and b and exploratory research question)

Gender. The negative effect of gender on CPS performance was mediated by the number of interactions. The indirect effect was negative and significant (Model 1: indirect effect=-0.19, p<.001) and exhibited significant between-country variation (Model 1: τ(indirect effect)=0.07, p<.001), meaning that boys presumably exhibited stronger CPS performance due to performing a larger number of interactions. The direct effect was also significant (Model 1: direct effect =-0.08, p<.001) and did not exhibit between-country variation (Model 1: τ(direct effect)=0.04, p=.241). This meant that while overall boys outperformed girls in CPS, *conditional on a specific number of interactions*, the performance difference was clearly smaller. The median κ^2 was 0.13, which reflects a medium to large sized mediation effect (Cohen, 1988), supporting Hypothesis 4b.

The amount of exploration seemed to be an even stronger mediator. The indirect effect was negative (Model 2: indirect effect =-0.23, p=.077) and did not vary between countries (Model 2: τ(indirect effect)=0.03, p=.382). This meant that the better performance of boys as compared to girls was not only related to the larger amount of interaction overall, but specifically related to a larger amount of exploration. The direct effect was not significant (Model 2: direct effect =-0.03, p=.083). This meant that after controlling for exploration, girls and boys showed equal CPS performance. The direct effect also showed no between-country variation (Model 2:

τ(direct effect)=0.05, p=.292). The median κ^2 was 0.17, which reflects a medium to large sized effect according to Cohen's conventions (1988), and indicates a larger effect than in Model 1.

Migration Background. The results from Model 3 indicated no mediation of the effect of migration background on CPS performance by the number of interactions. The indirect effect was not significant (Model 3: indirect effect=-0.07, p=.077). It did, however, vary between countries (Model 3: τ(indirect effect)=0.19, p<.001). The direct effect was not significant either, (Model 3: direct effect=-0.07, p=.146; total effect=-0.16, p=.001) and also showed between-country variation (Model 3: τ(direct effect)=0.27, p<.001). The direct effect not being significant meant that after controlling for the number of interactions, and reading skill, students without a migration background did not outperform students with a migration background.

The effect of migration background on CPS performance was not mediated by the amount of exploration. The indirect effect was not significant (Model 4: indirect effect=0.02, p=.235). It did, however, vary between countries (Model 4: τ(indirect effect)=0.08, p<.001). The direct effect was of similar size as the total effect and significant (Model 4: direct effect=-0.15, p<.001; total effect=-0.12, p=.003). The direct effect also varied between countries (Model 4: τ(direct effect)=0.15, p<.001). Therefore, we conclude that performance differences in CPS between students with and without a migration background cannot be explained by the amount of interaction or the amount of exploration.

Discussion

In the present study, we used process measures to explain performance differences in CPS between groups. We used data from over 81,000 students from 42 countries. Therefore, those results that did not show between-country variation can be generalized across many different cultures. We first investigated performance differences in CPS between girls and boys

and between students with and without a migration background. Second, we investigated how interactive and explorative processes are related to CPS performance. Third, we investigated whether students of different gender or migration status differed in their interactive and explorative behavior. Finally, we tested whether the observed performance differences can be explained by differences in interactive or explorative behavior. In the following section, we will discuss our results.

Performance differences between groups

Our results indicate that boys exhibit higher performance in CPS than girls. In PISA 2012, boys exhibited higher problem solving performance than girls (OECD, 2014b). Our finding extends the results of PISA 2012 to the subdomain CPS. The results also show that students with a migration background exhibited lower performance than students without a migration background. Again, this extends PISA 2012 findings in which students with a migration background exhibited lower performance in problem solving (OECD, 2014b). Notably, both group differences showed between-country variation. In some countries, the observed group differences were not present at all or were even reversed. For instance, in the United Arab Emirates, girls exhibited higher CPS performance than boys, and in Singapore, students with a migration background exhibited higher performance than students without a migration background. Therefore, when referring to specific countries, the specific findings and patterns of those countries should be considered as well. We will discuss possible causes for the observed performance differences in the section “Mediation of performance differences”.

Effects of behavior indicators on CPS performance

Both of our process measures (number of interactions and number of exploration steps) were positively related to CPS performance. This finding is in line with the results of Naumann

et al. (2014), who showed that less interaction is detrimental for problem solving performance in technology-based environments. We also extended the findings of Bell and Kozlowski (2008) and Dormann and Frese (1994) concerning exploration in error training to CPS. However, the relation between interaction and CPS was stronger than the relation between exploration and CPS. This indicates that goal-directed behavior is more important than exploration for solving a complex problem. This finding is not surprising, since a complex problem cannot be solved without goal-directed behavior; meanwhile, exploration should facilitate the execution of goal-directed behavior but is not immediately necessary to solve a problem. These effects also showed between-country variation. However, both effects were positive in all countries, making a strong argument for a general positive relation between interaction/exploration and CPS performance.

Behavioral differences between groups

As expected, we found behavioral differences between boys and girls. Boys exhibited more interactive behavior and also more exploration. These results extend the findings of Wittmann and Hattrup (2004), who found that boys exhibit more risky behavior, and Wüstenberg et al. (2014), who showed that boys used the VOTAT strategy more often. Both risky behavior as described by Wittmann and Hattrup (2004) and the VOTAT strategy as described by Wüstenberg et al. (2014) can be regarded as special cases of exploration behavior. The risky behavior in the study by Wittmann and Hattrup (2004) involved the use of extreme values in a simulated problem scenario. On the one hand, the use of these extreme input values carried the risk of causing "catastrophic" effects in the simulation. On the other hand, causing more dramatic changes in the simulation by choosing extreme input values provided more information about the structure of the problem than moderate values would have produced. Therefore, this behavior can be regarded as exploration. The VOTAT strategy described by Wüstenberg et al. (2014) can also

be seen as a special case of exploration behavior. The VOTAT strategy is applicable to items in the linear structural equation systems framework (Greiff & Funke, 2009). It implies that every input variable should be manipulated separately to find out which relations between input and output variables exist. Students generate more knowledge about the problem structure by applying this strategy more frequently (Wüstenberg et al., 2014), which is why it might also be regarded as exploration behavior. Therefore, our findings can be regarded as a generalization of these previous results. Although we found between-country variation, the estimates in all countries for the number of exploration steps and in the majority of countries for the number of interactions pointed in the same direction. An exception was again the United Arab Emirates, where we found that females engaged in more interactions than males. These differences in the relation of gender and behavior between countries could be related to the respective culture and its predominant gender norms. Culture-specific gender norms could influence the behavior of girls and boys in different ways, including differences in education for girls and boys both at school and at home.

We found no relation between migration background and number of interactions, indicating that students with a migration background interacted as much with the problems as students without a migration background. This finding does not support the results of Sonnleitner et al. (2014), who found that immigrant students in Luxembourg interacted more with the problems than their peers without a migration background. However, the immigrant population in Luxembourg may have particular characteristics that do not occur in other countries. We will further discuss this issue in the section “Mediation of performance differences”. We did not find any effect of migration background on the amount of exploration. Therefore, Sonnleitner et al.’s argumentation (2014) that immigrant students do not have a deficit in exploration behavior but

have difficulties utilizing the generated information can be applied to our results. We also found significant between-country variation and a substantial number of countries with reversed estimates for both effects. Therefore, these effects should be interpreted in a country-conditional manner.

Mediation of performance differences

Our results showed that the performance difference between boys and girls was mediated by both process measures, with amount of exploration being a stronger mediator than the number of interactions, despite the fact that the number of interactions was the stronger predictor of CPS performance. In other words, a lack of interaction and even more so a lack of exploration prevent girls from exhibiting equally high CPS performance as boys. One reason why girls exhibit less interaction and exploration might be that girls are encouraged less often to engage in this kind of behavior than boys. For example, Cherney and London (2006) argue that play with different kinds of toys may foster the development of different cognitive abilities in girls and boys. They found that boys between 5 and 13 years of age preferred toys that encourage manipulation, construction, and exploration, while girls in the same age group preferred toys that encourage the development of verbal skills. Leaper and Friedman (2007) argue that children start even earlier to develop gender-related cognitions. They state that three-year-old children are already aware of their own gender-group membership which becomes part of their social identity. Between 3 and 6 years of age, children begin to form stereotypes about gender-specific activities. Therefore, girls might not be motivated or may not even come up with the idea of exhibiting exploration behavior and therefore exhibit lower CPS performance. Miller (1987) also argues that girls are discouraged to solve problems by certain socialization practices such as the discouragement of active play and the restriction of exploration through parents, teachers, peers, the media, and

cultural institutions. Therefore, these processes could be promising starting points for improving girls' CPS performance. This difference in the socialization of girls and boys may differ between cultures, since we found between-country variation of the relation of gender and interaction/exploration behavior. Another reason why girls might be less motivated to engage in interaction and exploration behavior could be lower self-efficacy in computer-based environments. For example, one finding of PISA 2012 was that in most countries boys are exposed to computers much earlier than girls (OECD, 2015b). The OECD (2015b) argues that restricting girls' access to computers might lower their self-efficacy in computer-based tasks. The OECD (2015c) also found that on average girls spend less of their leisure time engaging with computers and that they less often have career ambitions in the field of computing and engineering than boys. Spending less time with computers girls might indeed develop a lower computer-related self-efficacy than boys. Again, these relations may differ between countries, which is reflected by the between-country variation of the relation between gender and behavior we found. However, it seems implausible that differences in computer skills caused the observed CPS performance differences since only basic computer skills were needed to solve the items and a tutorial was used to make sure that all participants were able to operate the computer-based testing environment. Moreover, the OECD (2016) found in their PIAAC study that basic computer skills are quite balanced across females and males and that poor basic computer skills are rarely found among young adults. Punter, Meelissen, and Glas (2017) did also not find gender differences in applying technical functionality using data from the International Computer and Information Literacy Study (ICILS). Moreover, Greiff, Kretzschmar, Müller, Spinath, and Martin (2014) found only weak to moderate relations between CPS ability and computer skills in several studies. For these reasons, we think it unlikely that the CPS performance differences we

observed are a function of computer skills differing between genders. Although not lacking the required computer skills, assessing CPS using computer-based items could have discouraged girls to some extent (due to lower computer-related self-efficacy), leading them to exhibit less interactive behavior and thus lower performance than boys. Again, these effects show between-country variation, so when referring only to a single country its specific effect should be considered. For instance, in Estonia no mediation of the gender differences by the number of interactions was observed. If indeed gender performance differences result from gender-specific childhood experiences regarding the encouragement of behavior and access to technology, the differences in CPS between countries might be partly a result of cultural differences in gender norms and socialization. This socialization could also lead to lower technology-related self-efficacy or higher anxiety among girls which in turn leads more passive behavior as Naumann et al. (2014) argue. However, interaction and exploration can also be subject to educational interventions in the respective countries. In the United Arab Emirates in which we found girls to show more interactions the educational system might promote this kind of behavior (see Appendix B).

We found that the performance difference between students with and without a migration background was neither mediated by the number of interactions nor by the amount of exploration. Sonnleitner et al. (2014) argue that immigrant students engage in more exploration than students without a migration background. However, they state that immigrant students have difficulties transferring the generated information into declarative knowledge and therefore are outperformed in CPS by students without a migration background. Our results support the view that immigrant students exhibit exploration behavior equal to non-immigrant students. Although exhibiting exploration behavior they might not be able to process the required information. One

possible reason why students with migration background might not profit from exploration behavior (see Table 2) might be a low proficiency in the test language. However, since we controlled for reading ability in our models, neither a low language ability nor a lack of exploration behavior seem to be the primary cause for the low CPS performance of immigrant students. Sonnleitner et al. (2014) investigated the relation between CPS and migration background. They found that the lower performance of immigrant students could be explained by students with migration background being much more often enrolled in lower academic tracks. Also Greiff et al. (2013) found a strong relation between students' CPS ability and their academic achievement.

Theoretically, the country-specific mechanisms leading to lower CPS performance among students with a migration background could be an effect of the composition of the migrant population in each country. For example, in countries in which most immigrants received little education, these deficits in education could be the main reason for immigrants' underachievement (e.g. countries mostly recruiting workers for rather simple jobs from abroad, leading to immigration of low educated people). However, in countries in which most immigrants are highly educated, there could be other reasons for performance differences (e.g. countries recruiting mostly highly educated people from abroad). Another factor that might be related to the CPS performance of students with a migration background is their socio-economic status. In PISA 2012, socio-economic status (SES) was related to many competencies, including problem solving (OECD, 2014b). Therefore, the socio-economic composition of the immigrant population in each country could also affect their CPS performance. In this case, it would not be appropriate to generalize this effect across countries. The SES could also be related to students' access to technology, since low SES households might not be able to afford computers or

laptops. Therefore, lower computer abilities among low SES students might be a confounding variable to the computer-based CPS assessment. However, as stated before the tasks only required very rudimental computer skills, and a tutorial was offered, so that every student should have been able to solve the tasks as far as operating the technological interface is concerned. Moreover, Greiff et al. (2014) found only weak to moderate relations between CPS and computer skills. Future research however might further address these questions by examining the socio-economic status and level of education of migrants in different countries. Interaction effects between gender and migration that might vary between countries would also be conceivable and might be addressed in future research. Moreover, not only the country of assessment but also the countries where migrants originate from might play a role in this regard.

Limitations

Since our data refers only to fifteen-year-old students, the generalizability of our results to different age groups is limited. Like all studies that use PISA data we cannot rule out possible confounding variables like computer skills in our analyses. Furthermore, we had to exclude the data from Korea since not all models converged for this country's dataset. Another limitation of our study is the fact that we had to exclude several items. For future research, it would be preferable to have more complete and unambiguous log data to avoid item drop-outs. We also had to exclude some items due to a restriction in exploration. The exclusion of these items might have led to an over- or underrepresentation of certain item characteristics in the resulting item pool. It should also be kept in mind that we did not use any experimental manipulation. Therefore, our results are of a correlational nature and should be further confirmed with experimental designs in future research. Moreover, we only used two somewhat arbitrary behavioral indicators to represent interactive and explorative behavior. Especially with regard to

exploration, different operationalizations would also be possible. Since we defined all interactions that were not necessary for item completion as exploration steps, our definition was rather broad. However, this broad definition was necessary to align with the heterogeneous items in the PISA assessment. Another limitation of our findings is that we cannot explain the variation in the observed effects between countries. Future research should try to explain this variation by examining country characteristics, for instance differences in the composition of immigrant groups. But also differences in the respective curricula could lead to country-specific effects. Especially, CPS being part of a country's curriculum could heavily influence the results. Furthermore, a country's error culture could influence the extent to which students are willing to explore which might lead to errors. This between-country variation also limits the generalizability of our results making it necessary to take into account the context of the countries of interest when interpreting the results.

The differences between the effects of gender and migration background between countries might also be related to cultural differences in gender roles or differences in the migration population. For example Naumann, Elson, and Rauch (2016, April) found that in digital reading, a domain with close ties to complex problem solving (Brand-Gruwel, Wopereis, & Walraven, 2009), there were stark differences between economies in the effect of migration background on task-adaptive navigation. These effects were strong in European, but absent or weak in Oceanic and Chinese economies, where Chinese students, who generally perform very well abroad (OECD, 2015a), make a large part of the immigrant population. A more detailed analysis of this issue however must be left to future research, using data sets where explicit information on the countries immigrant students migrated from is available.

In a similar vein, future research might look at interactive effects between gender and migration background, conditional on the culture from where immigrant students came, on CPS behavior and performance. It might well be the case that disadvantages for girls and for students with migration background overlap and reinforce each other.

Moreover, not differentiating between first and second generation migration background might have had weakened the effects found with respect to migration background. The OECD (2014) reported an even lower problem solving performance of students who were born outside the country of assessment than students who were born in the country of assessment (but whose parents were born abroad). Therefore, in future research these groups could be investigated separately to reveal differential effects.

Conclusion

Our results indicated that the performance difference in CPS between boys and girls can be explained by interaction and exploration behavior. Since exploration is part of overall interactive behavior and exploration more strongly mediates the gender effect, explorative behavior might just be the crucial factor causing performance differences between girls and boys in CPS. As soon as interaction or exploration was taken into account, girls' performance was equal to boys' performance. Therefore, girls' lower performance in CPS might be due to a lower readiness to explore. On the other hand, the performance difference between students with and without a migration background cannot be explained by explorative or interactive behavior. Notably, in some countries, students with a migration background exhibited more interactions than students without a migration background, while in other countries, this effect was reversed. Nevertheless, students with a migration background exhibited lower performance in most

countries. Therefore, behavioral differences do not seem to be the primary cause for the lower CPS performance of students with a migration background.

In our study, we could show that measures derived from log data can serve as mediating variables to explain performance differences between groups and thus shed light on the mechanisms behind these performance differences. We could show that gender differences in CPS performance seem to stem from gender-specific behavior, while differences by migration status seem to have other causes. Therefore, besides showing that log data can be used to predict performance, we also showed that performance differences between groups can be explained by behavioral differences as recorded in log data. This means that future studies using computer-based assessment data could report not only group differences in performance, but also possible causes of these differences.

References

Artelt, C., Naumann, J., & Schneider, W. (2010). Lesemotivation und Lernstrategien. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel,. . . P. Stanat (Eds.), *PISA 2009: Bilanz nach einem Jahrzehnt* (pp. 73–112). Münster: Waxmann.

Bell, B. S., & Kozlowski, S. W. J. (2008). Active learning: effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology*, *93*, 296–316. https://doi.org/10.1037/0021-9010.93.2.296

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining Twenty-First Century Skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 17–66). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-2324-5_2

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education*, *53*, 1207–1217. https://doi.org/10.1016/j.compedu.2009.06.004

Cherney, I. D., & London, K. (2006). Gender-linked Differences in the Toys, Television Shows, Computer Games, and Outdoor Activities of 5- to 13-year-old Children. *Sex Roles*, *54*, 717–726. https://doi.org/10.1007/s11199-006-9037-8

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cross, C. P., Copping, L. T., & Campbell, A. (2011). Sex differences in impulsivity: a meta-analysis. *Psychological Bulletin*, *137*, 97–130. https://doi.org/10.1037/a0021591

Dormann, T., & Frese, M. (1994). Error training: Replication and the function of exploratory behavior. *International Journal of Human-Computer Interaction*, *6*, 365–372. https://doi.org/10.1080/10447319409526101

Frensch, P. A., & Funke, J. (1995). Definitions, Traditions, and a General Framework for Understanding Complex Problem Solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 3–22). Hillsdale, NJ: L. Erlbaum Associates.

Frese, M., & Keith, N. (2015). Action errors, error management, and learning in organizations. *Annual Review of Psychology*, *66*, 661–687. https://doi.org/10.1146/annurev-psych-010814-015205

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*, 608–626. https://doi.org/10.1037/a0034716

Greiff, S., & Funke, J. (2009). Measuring Complex Problem Solving: The MicroDYN approach. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 157–163).

Greiff, S., Kretzschmar, A., Müller, J. C., Spinath, B., & Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *Journal of Educational Psychology*, *106*, 666–680. https://doi.org/10.1037/a0035426

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105.

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, *105*, 364–379. https://doi.org/10.1037/a0031856

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–18. https://doi.org/10.1080/10705511.2017.1402334

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, *26*, 379–424. https://doi.org/10.1080/07370000802212669

Keith, N., & Frese, M. (2005). Self-regulation in error management training: Emotion control and metacognition as mediators of performance effects. *Journal of Applied Psychology*, *90*, 677–691. https://doi.org/10.1037/0021-9010.90.4.677

Keith, N., & Frese, M. (2008). Effectiveness of error management training: a meta-analysis. *Journal of Applied Psychology*, *93*, 59–69. https://doi.org/10.1037/0021-9010.93.1.59

Keith, N., Richter, T., & Naumann, J. (2010). Active/Exploratory training promotes transfer even in learners with low motivation and cognitive ability. *Applied Psychology*, *59*, 97–123. https://doi.org/10.1111/j.1464-0597.2009.00417.x

Kelley, K., & Lai, K. (2010). MBESS: MBESS. R package. Retrieved from https://CRAN.R-project.org/package=MBESS

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log-data from technology-based assessments? - A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*, 527–563.

Leaper, C., & Friedman, C. K. (2007). The Socialization of Gender. In J. E. Grusec & P. D. Hastings (Eds.), *Handbook of Socialization: Theory and Research* (pp. 561–587). New York: Guilford Publications.

Martin, A. J., Liem, G. A. D., Mok, M. M. C., & Xu, J. (2012). Problem solving and immigrant student mathematics and science achievement: Multination findings from the Programme for International Student Assessment (PISA). *Journal of Educational Psychology*, *104*, 1054–1073. https://doi.org/10.1037/a0029152

Mayer, R. E., & Wittrock, M. C. (2006). Problem Solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 287–304). Mahwah, N.J: Erlbaum.

Miller, C. L. (1987). Qualitative Differences Among Gender-Stereotyped Toys: Implications for Cognitive and Social Development in Girls and Boys. *Sex Roles*, *16*, 473–487.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

Naumann, J. (2015). A model of online reading engagement: Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior*, *53*, 263–277. https://doi.org/10.1016/j.chb.2015.06.051

Naumann, J. (2019). The Skilled, the Knowledgeable, and the Motivated: Investigating the Strategic Allocation of Time on Task in a Computer-Based Assessment. *Frontiers in psychology*, *10*.

Naumann, J., Elson, M., & Rauch, D. (2016, April). *Explaining performance gaps in digital reading between native and immigrant students through group-specific navigation behavior*. AERA Annual Meeting, Washington, D.C.

Naumann, J., Goldhammer, F., Rölke, H., & Stelter, A. (2014). Erfolgreiches Problemlösen in technologiebasierten Umgebungen: Wechselwirkungen zwischen Interaktionsschritten und Aufgabenanforderungen. *Zeitschrift für Pädagogische Psychologie*, *28*, 193–203. https://doi.org/10.1024/1010-0652/a000134

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, *33*, 31–88. https://doi.org/10.1145/375360.375365

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in germany. *International Journal of Science and Mathematics Education*, *8*, 545–563. https://doi.org/10.1007/s10763-010-9206-7

OECD. (2011). *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)* (1. Aufl.). *PISA 2009 Results*: OECD. Retrieved from http://gbv.eblib.com/patron/FullRecord.aspx?p=655759

OECD. (2013a). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. *PISA*. Paris: OECD.

OECD. (2014a). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*. Paris: OECD Publishing. Retrieved from http://gbv.eblib.com/patron/FullRecord.aspx?p=1683156

OECD. (2014b). *PISA 2012 Results: Creative Problem Solving: Students' skills in tackling real-life problems* (Volume V). Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264208070-en

OECD. (2014c). *PISA 2012: Technical Report*. Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD. (2015a). *Immigrant Students at School: Easing the Journey towards Integration*. *OECD Reviews of Migrant Education*. Paris: OECD Publishing.

OECD. (2015b). *Students, computers and learning: Making the connection*. Paris: OECD Publishing.

OECD. (2015c). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*. *PISA*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264229945-en

OECD. (2016). *Skills matter: Further results from the survey of adult skills*. *OECD skills studies*. Paris: OECD Publishing.

OECD. (2013b). *OECD skills outlook 2013*. Paris: OECD Publishing.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*, 93–115. https://doi.org/10.1037/a0022658

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M.,. . . Schiefele, U. (Eds.). (2004). *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland : Ergebnisse des zweiten internationalen Vergleichs*. Münster, New York: Waxmann.

Punter, R. A., Meelissen, M. R. M., & Glas, C. A. W. (2017). Gender differences in computer and information literacy: An exploration of the performances of girls and boys in ICILS 2013. *European Educational Research Journal*, *16*, 762–780.

R Core Team. (2016). *R: A language and environment for statistical computing: R Foundation for Statistical Computing*. Vienna. Retrieved from https://www.R-project.org/

Sonnleitner, P., Brunner, M., Keller, U., & Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *Journal of Educational Psychology*, *106*, 681–695. https://doi.org/10.1037/a0035506

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Wittmann, W. W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science*, *21*, 393–409. https://doi.org/10.1002/sres.653

Wüstenberg, S., Greiff, S., Molnár, G., & Funke, J. (2014). Cross-national gender differences in complex problem solving and their determinants. *Learning and Individual Differences*, *29*, 18–29. https://doi.org/10.1016/j.lindif.2013.10.006

Appendix A

Appendix A. Sample size, percentage of females and percentage of students with a migration background by country.

Country	n	% female	% with migration background
Australia	5608	49.34	19.08
Austria	1328	50.98	16.64
Belgium	2145	49.84	14.08
Brazil	1455	50.03	1.03
Bulgaria	2138	48.97	0.65
Canada	4584	50.22	18.08
Chinese Taipei	1483	52.06	0.74
Colombia	2286	54.07	0.22
Croatia	1923	50.81	12.01
Czech Republic	3076	50.26	3.38
Denmark	1948	52.67	24.64
Estonia	1363	51.14	7.85
Finland	3531	48.17	12.29
France	1344	51.93	13.54
Germany	1350	48.52	9.93
Hong Kong-China	1323	46.64	31.67
Hungary	1300	51.92	1.77
Ireland	1188	51.43	9.43
Israel	1341	56.67	17.52
Italy	1370	45.11	7.88
Japan	3011	48.12	0.33
Macao-China	1564	49.94	59.40
Malaysia	1927	51.32	1.50
Montenegro	1845	52.57	6.23
Netherlands	1752	48.92	10.62
Norway	1237	48.42	9.46
Poland	1227	50.29	0.16
Portugal	1444	50.21	7.55
Russian Federation	1537	49.06	9.82
Serbia	1775	51.21	8.62
Shanghai-China	1203	51.29	0.91
Singapore	1392	49.28	16.95
Slovak Republic	1463	45.80	0.68
Slovenia	2064	45.06	9.45
Spain	2703	50.13	8.95
Sweden	1256	52.15	15.29
Switzerland	1575	52.25	1.40

Turkey	1995	48.12	0.80
United Arab Emirates	3246	50.89	52.34
United Kingdom	1456	52.95	13.26
United States of America	1271	50.83	18.96
Uruguay	2012	52.09	0.60

Appendix B

Appendix B: Model 1 results and effect size κ^2 by country with gender as predictor and number of interactions as mediator

Country	a	SD	b	SD	direct	SD	indirect	SD	total	SD	κ^2
Australia	-0.18	0.05	0.66	0.03	-0.06	0.05	-0.12	0.03	-0.18	0.04	0.06
Austria	-0.39	0.10	0.66	0.05	-0.22	0.09	-0.26	0.07	-0.48	0.08	0.19
Belgium	-0.19	0.06	0.72	0.03	-0.07	0.06	-0.14	0.05	-0.21	0.06	0.09
Brazil	-0.28	0.10	0.75	0.07	-0.19	0.10	-0.21	0.08	-0.40	0.10	0.13
Bulgaria	-0.18	0.06	0.70	0.03	-0.18	0.06	-0.13	0.04	-0.31	0.06	0.05
Canada	-0.24	0.07	0.61	0.04	-0.04	0.06	-0.14	0.04	-0.18	0.06	0.08
Chinese Taipei	-0.37	0.09	0.62	0.06	-0.09	0.09	-0.23	0.06	-0.32	0.08	0.18
Colombia	-0.26	0.08	0.71	0.06	-0.26	0.10	-0.18	0.06	-0.45	0.09	0.13
Croatia	-0.33	0.06	0.76	0.03	-0.19	0.05	-0.25	0.05	-0.44	0.06	0.14
Czech Republic	-0.40	0.07	0.65	0.04	0.03	0.05	-0.26	0.04	-0.23	0.05	0.20
Denmark	-0.21	0.09	0.72	0.07	-0.30	0.10	-0.15	0.06	-0.45	0.10	0.05
Estonia	-0.16	0.10	0.73	0.06	-0.17	0.10	-0.12	0.07	-0.29	0.10	0.00
Finland	-0.35	0.06	0.70	0.04	0.04	0.06	-0.25	0.05	-0.21	0.06	0.10
France	-0.24	0.08	0.69	0.05	-0.11	0.07	-0.17	0.06	-0.28	0.08	0.09
Germany	-0.14	0.08	0.74	0.04	-0.15	0.07	-0.10	0.06	-0.25	0.08	0.00
Hong Kong-China	-0.30	0.08	0.74	0.05	-0.10	0.08	-0.22	0.06	-0.32	0.09	0.17
Hungary	-0.27	0.08	0.69	0.06	-0.01	0.09	-0.19	0.06	-0.19	0.09	0.08
Ireland	-0.03	0.11	0.68	0.06	-0.16	0.11	-0.02	0.08	-0.18	0.11	0.03
Israel	-0.25	0.09	0.78	0.04	-0.16	0.07	-0.20	0.07	-0.36	0.09	0.11
Italy	-0.38	0.14	0.83	0.04	-0.04	0.11	-0.32	0.12	-0.35	0.11	0.30
Japan	-0.38	0.06	0.67	0.04	-0.07	0.07	-0.25	0.05	-0.32	0.07	0.20
Macao-China	-0.48	0.10	0.66	0.07	-0.01	0.12	-0.32	0.08	-0.33	0.12	0.21
Malaysia	-0.19	0.07	0.67	0.04	0.00	0.06	-0.13	0.04	-0.13	0.06	0.06
Montenegro	-0.15	0.08	0.80	0.03	0.03	0.09	-0.12	0.06	-0.09	0.11	0.06
Netherlands	-0.04	0.09	0.80	0.04	0.00	0.06	-0.03	0.07	-0.03	0.08	0.07
Norway	-0.22	0.10	0.71	0.06	-0.03	0.09	-0.16	0.07	-0.19	0.09	0.08
Poland	-0.34	0.10	0.74	0.05	0.04	0.10	-0.25	0.08	-0.22	0.08	0.20
Portugal	-0.30	0.10	0.76	0.05	-0.14	0.09	-0.23	0.08	-0.36	0.10	0.18
Russian Federation	-0.24	0.09	0.66	0.05	-0.08	0.08	-0.16	0.06	-0.24	0.09	0.07
Serbia	-0.35	0.07	0.70	0.06	-0.14	0.06	-0.25	0.05	-0.39	0.07	0.14
Shanghai-China	-0.53	0.09	0.55	0.06	-0.01	0.10	-0.29	0.06	-0.31	0.09	0.20
Singapore	-0.47	0.08	0.56	0.06	0.01	0.09	-0.27	0.05	-0.26	0.08	0.18

Slovak Republic	-0.38	0.09	0.60	0.05	-0.17	0.08	-0.23	0.06	-0.40	0.08	0.16
Slovenia	-0.06	0.09	0.81	0.05	-0.12	0.08	-0.05	0.07	-0.17	0.09	0.07
Spain	-0.23	0.10	0.86	0.04	-0.05	0.10	-0.20	0.09	-0.25	0.10	0.18
Sweden	-0.20	0.09	0.80	0.04	-0.16	0.08	-0.16	0.07	-0.32	0.09	0.02
Switzerland	-0.45	0.11	0.76	0.06	-0.10	0.11	-0.34	0.09	-0.45	0.09	0.26
Turkey	-0.59	0.06	0.62	0.04	-0.10	0.08	-0.36	0.05	-0.46	0.07	0.22
United Arab Emirates	0.12	0.07	0.73	0.04	-0.02	0.06	0.08	0.05	0.06	0.07	0.20
United Kingdom	-0.35	0.09	0.62	0.08	-0.20	0.12	-0.22	0.06	-0.42	0.12	0.16
United States of America	-0.35	0.13	0.57	0.07	-0.02	0.10	-0.20	0.08	-0.22	0.10	0.13
Uruguay	-0.20	0.06	0.78	0.03	-0.03	0.06	-0.16	0.05	-0.18	0.07	0.09

Note. a: effect of gender on interaction. b: effect of interaction on CPS.

Appendix C

Appendix C: Model 2 results and effect size κ^2 by country with gender as predictor and amount of exploration as mediator.

Country	a	SD	b	SD	direct	SD	indirect	SD	total	SD	κ^2
Australia	-0.56	0.06	0.45	0.07	0.07	0.06	-0.25	0.04	-0.18	0.05	0.18
Austria	-0.67	0.11	0.49	0.10	-0.15	0.12	-0.33	0.09	-0.48	0.08	0.25
Belgium	-0.53	0.10	0.52	0.07	0.05	0.08	-0.28	0.06	-0.23	0.06	0.19
Brazil	-0.28	0.24	0.36	0.10	-0.30	0.14	-0.10	0.09	-0.40	0.10	0.06
Bulgaria	-0.52	0.09	0.47	0.10	-0.06	0.09	-0.25	0.07	-0.31	0.07	0.13
Canada	-0.51	0.09	0.33	0.10	-0.02	0.08	-0.17	0.05	-0.18	0.06	0.11
Chinese Taipei	-0.64	0.10	0.56	0.09	0.03	0.11	-0.36	0.10	-0.33	0.08	0.28
Colombia	-0.50	0.15	0.70	0.13	-0.14	0.16	-0.35	0.13	-0.49	0.09	0.27
Croatia	-0.53	0.08	0.50	0.06	-0.18	0.08	-0.26	0.05	-0.44	0.06	0.17
Czech Republic	-0.65	0.07	0.42	0.07	0.04	0.07	-0.28	0.05	-0.23	0.05	0.21
Denmark	-0.48	0.20	0.60	0.65	-0.17	0.40	-0.29	0.38	-0.46	0.10	0.18
Estonia	-0.42	0.12	0.53	0.08	-0.08	0.11	-0.22	0.07	-0.30	0.10	0.11
Finland	-0.67	0.07	0.28	0.11	-0.01	0.11	-0.19	0.08	-0.20	0.06	0.12
France	-0.50	0.10	0.39	0.08	-0.10	0.09	-0.19	0.05	-0.30	0.08	0.12
Germany	-0.50	0.14	0.45	0.13	-0.05	0.12	-0.23	0.10	-0.28	0.08	0.16
Hong Kong-China	-0.57	0.10	0.50	0.09	-0.04	0.10	-0.28	0.08	-0.32	0.10	0.19
Hungary	-0.47	0.16	0.38	0.12	-0.02	0.11	-0.18	0.09	-0.20	0.10	0.13
Ireland	-0.51	0.17	0.33	0.18	-0.01	0.14	-0.17	0.10	-0.18	0.11	0.11
Israel	-0.39	0.13	0.52	0.10	-0.15	0.10	-0.21	0.07	-0.36	0.08	0.12
Italy	-0.80	0.22	0.48	0.18	0.01	0.22	-0.38	0.21	-0.37	0.11	0.24
Japan	-0.61	0.06	0.45	0.06	-0.05	0.07	-0.28	0.05	-0.32	0.07	0.17
Macao-China	-0.70	0.11	0.44	0.08	-0.01	0.12	-0.31	0.07	-0.33	0.12	0.18
Malaysia	-0.53	0.08	0.46	0.07	0.11	0.08	-0.24	0.05	-0.13	0.07	0.17
Montenegro	-0.41	0.08	0.49	0.08	0.10	0.11	-0.20	0.05	-0.10	0.11	0.10
Netherlands	-0.23	0.14	0.41	0.16	0.06	0.08	-0.09	0.04	-0.04	0.07	0.07
Norway	-0.72	0.10	0.36	0.11	0.09	0.12	-0.26	0.09	-0.17	0.09	0.19
Poland	-1.01	0.14	0.57	0.20	0.36	0.26	-0.58	0.25	-0.22	0.08	0.39
Portugal	-0.48	0.10	0.56	0.09	-0.11	0.11	-0.27	0.07	-0.38	0.10	0.18
Russian Federation	-0.69	0.10	0.26	0.08	-0.06	0.11	-0.18	0.06	-0.24	0.09	0.11
Serbia	-0.52	0.08	0.44	0.07	-0.15	0.08	-0.23	0.05	-0.38	0.07	0.14
Shanghai-China	-0.75	0.14	0.35	0.12	-0.05	0.13	-0.26	0.10	-0.31	0.09	0.16
Singapore	-0.65	0.09	0.37	0.09	0.00	0.11	-0.24	0.07	-0.25	0.08	0.16
Slovak Republic	-0.82	0.10	0.32	0.10	-0.13	0.12	-0.26	0.08	-0.39	0.08	0.20

Slovenia	-0.50	0.14	0.79	0.15	0.20	0.13	-0.40	0.12	-0.20	0.09	0.32
Spain	-0.55	0.12	0.53	0.12	0.04	0.12	-0.29	0.09	-0.25	0.11	0.18
Sweden	-0.55	0.13	0.42	0.13	-0.13	0.11	-0.23	0.08	-0.36	0.09	0.14
Switzerland	-0.69	0.12	0.49	0.08	-0.12	0.13	-0.34	0.09	-0.45	0.09	0.19
Turkey	-0.67	0.08	0.33	0.08	-0.24	0.09	-0.22	0.06	-0.46	0.07	0.15
United Arab Emirates	-0.18	0.10	0.56	0.06	0.15	0.08	-0.10	0.06	0.05	0.07	0.00
United Kingdom	-0.74	0.08	0.48	0.09	-0.08	0.13	-0.35	0.08	-0.43	0.12	0.20
United States of America	-0.70	0.13	0.28	0.08	-0.02	0.10	-0.20	0.07	-0.22	0.10	0.12
Uruguay	-0.43	0.11	0.33	0.08	-0.05	0.08	-0.14	0.06	-0.19	0.07	0.09

Note. a: effect of gender on exploration. b: effect of exploration on CPS.

Appendix D

Appendix D: Model 3 results and effect size κ^2 by country with migration background as predictor and number of interactions as mediator.

Country	a	SD	b	SD	direct	SD	indirect	SD	total	SD	κ^2
Australia	0.23	0.06	0.67	0.03	-0.09	0.06	0.15	0.04	0.07	0.06	0.07
Austria	-0.19	0.13	0.69	0.05	-0.27	0.13	-0.13	0.09	-0.41	0.13	0.02
Belgium	-0.34	0.11	0.77	0.03	-0.19	0.10	-0.26	0.09	-0.45	0.11	0.05
Brazil	-1.27	0.32	0.77	0.07	-0.22	0.76	-0.98	0.25	-1.20	0.75	0.10
Bulgaria	0.16	0.34	0.76	0.03	0.25	0.82	0.12	0.26	0.37	0.90	0.01
Canada	0.15	0.11	0.64	0.04	-0.34	0.08	0.09	0.07	-0.25	0.09	0.07
Chinese Taipei	0.70	0.20	0.65	0.06	-0.58	0.25	0.46	0.14	-0.12	0.30	NA
Colombia	-0.81	1.19	0.76	0.06	-0.70	2.46	-0.61	0.90	-1.31	1.76	0.03
Croatia	-0.10	0.10	0.81	0.04	-0.03	0.10	-0.08	0.08	-0.11	0.11	0.01
Czech Republic	-0.33	0.19	0.65	0.04	0.13	0.15	-0.21	0.13	-0.08	0.18	0.01
Denmark	-0.41	0.13	0.73	0.07	-0.16	0.12	-0.30	0.10	-0.46	0.13	0.05
Estonia	-0.01	0.15	0.76	0.06	-0.17	0.14	-0.01	0.11	-0.18	0.16	0.03
Finland	-0.53	0.18	0.72	0.04	0.12	0.13	-0.38	0.13	-0.25	0.10	0.00
France	-0.15	0.15	0.70	0.05	-0.26	0.14	-0.10	0.11	-0.37	0.16	0.03
Germany	-0.51	0.17	0.76	0.06	0.02	0.14	-0.39	0.13	-0.37	0.16	0.03
Hong Kong-China	0.10	0.10	0.78	0.05	-0.12	0.10	0.07	0.08	-0.04	0.11	0.00
Hungary	-0.04	0.30	0.71	0.07	-0.47	0.40	-0.03	0.21	-0.50	0.43	0.02
Ireland	0.06	0.19	0.72	0.06	-0.21	0.16	0.04	0.14	-0.17	0.17	0.04
Israel	0.03	0.12	0.82	0.04	0.20	0.09	0.03	0.10	0.22	0.13	0.03
Italy	-0.36	0.30	0.83	0.04	-0.23	0.30	-0.30	0.25	-0.53	0.31	0.01
Japan	0.75	0.52	0.68	0.04	-0.30	0.41	0.51	0.36	0.21	0.19	0.05
Macao-China	-0.08	0.08	0.67	0.08	-0.04	0.08	-0.05	0.06	-0.09	0.08	0.02
Malaysia	0.00	0.30	0.68	0.04	0.24	0.30	0.00	0.20	0.24	0.31	0.00
Montenegro	0.25	0.14	0.83	0.03	-0.10	0.10	0.21	0.12	0.12	0.14	0.01
Netherlands	-0.60	0.21	0.80	0.03	-0.27	0.10	-0.48	0.17	-0.74	0.20	0.01
Norway	-0.14	0.17	0.72	0.06	-0.46	0.18	-0.10	0.12	-0.56	0.16	0.03
Poland	-0.54	0.48	0.75	0.05	1.22	0.12	-0.40	0.36	0.82	0.40	NA
Portugal	0.08	0.16	0.81	0.05	-0.15	0.18	0.06	0.13	-0.09	0.18	0.04
Russian Federation	-0.18	0.17	0.67	0.05	-0.07	0.13	-0.12	0.11	-0.19	0.16	0.00
Serbia	-0.41	0.25	0.72	0.06	0.13	0.17	-0.30	0.17	-0.17	0.15	0.01
Shanghai-China	-1.38	0.72	0.58	0.06	-0.15	0.35	-0.80	0.42	-0.95	0.53	0.04
Singapore	0.06	0.11	0.59	0.06	0.04	0.11	0.04	0.06	0.08	0.11	0.03

Slovak Republic	-0.13	0.51	0.64	0.06	0.47	0.24	-0.08	0.33	0.38	0.30	0.02
Slovenia	-0.23	0.13	0.84	0.05	-0.04	0.19	-0.20	0.11	-0.23	0.18	0.10
Spain	-0.38	0.14	0.90	0.04	0.00	0.21	-0.34	0.13	-0.34	0.23	0.05
Sweden	-0.10	0.14	0.84	0.04	-0.17	0.13	-0.08	0.12	-0.26	0.15	0.03
Switzerland	-0.18	0.49	0.80	0.06	-0.32	0.25	-0.15	0.39	-0.46	0.34	0.02
Turkey	0.18	0.44	0.64	0.04	-0.14	0.28	0.11	0.28	-0.02	0.49	0.00
United Arab Emirates	0.49	0.07	0.74	0.04	0.07	0.07	0.36	0.05	0.43	0.07	0.15
United Kingdom	-0.14	0.13	0.62	0.07	-0.35	0.16	-0.09	0.08	-0.44	0.16	0.00
United States of America	0.34	0.13	0.59	0.07	-0.42	0.11	0.20	0.08	-0.22	0.10	0.05
Uruguay	0.13	0.36	0.79	0.03	0.25	0.43	0.10	0.29	0.35	0.35	0.02

Note. a: effect of migration on interaction. b: effect of interaction on CPS. NA indicates that κ^2 could not be calculated due to a zero covariance between migration background and interaction.

Appendix E

Appendix E: Model 4 results and effect size κ^2 by country with migration background as predictor and amount of exploration as mediator.

Country	a	SD	b	SD	direct	SD	indirect	SD	total	SD	κ^2
Australia	0.34	0.09	0.43	0.06	-0.10	0.06	0.15	0.04	0.05	0.06	0.07
Austria	0.05	0.16	0.51	0.06	-0.34	0.13	0.03	0.08	-0.32	0.12	0.02
Belgium	-0.24	0.13	0.52	0.06	-0.22	0.12	-0.12	0.07	-0.34	0.11	0.05
Brazil	-2.79	0.50	0.33	0.12	-0.24	0.75	-0.92	0.40	-1.16	0.69	0.10
Bulgaria	-0.22	0.41	0.41	0.07	0.57	0.88	-0.09	0.17	0.48	0.88	0.01
Canada	0.38	0.09	0.36	0.09	-0.38	0.09	0.14	0.05	-0.25	0.08	0.07
Chinese Taipei	0.00	0.40	0.50	0.08	-0.12	0.32	0.00	0.20	-0.12	0.31	NA
Colombia	-1.19	0.61	0.70	0.10	-0.04	2.23	-0.83	0.46	-0.87	1.89	0.03
Croatia	-0.08	0.12	0.48	0.06	-0.02	0.10	-0.04	0.06	-0.06	0.10	0.01
Czech Republic	-0.18	0.20	0.39	0.05	0.04	0.15	-0.07	0.08	-0.03	0.16	0.01
Denmark	-0.25	0.24	0.58	0.25	-0.25	0.20	-0.14	0.19	-0.39	0.13	0.05
Estonia	0.20	0.17	0.52	0.08	-0.21	0.17	0.10	0.09	-0.11	0.15	0.03
Finland	0.03	0.14	0.28	0.08	-0.11	0.08	0.01	0.04	-0.10	0.09	0.00
France	0.18	0.16	0.41	0.08	-0.38	0.14	0.08	0.07	-0.30	0.14	0.03
Germany	-0.24	0.23	0.39	0.17	-0.16	0.18	-0.09	0.11	-0.26	0.15	0.03
Hong Kong-China	0.02	0.10	0.48	0.08	-0.06	0.10	0.01	0.05	-0.06	0.10	0.00
Hungary	-0.53	0.72	0.32	0.50	-0.20	0.47	-0.17	0.46	-0.38	0.40	0.02
Ireland	0.28	0.23	0.37	0.12	-0.25	0.17	0.10	0.09	-0.14	0.15	0.04
Israel	0.16	0.15	0.48	0.08	0.16	0.11	0.08	0.07	0.23	0.12	0.03
Italy	0.02	0.30	0.51	0.16	-0.52	0.26	0.01	0.15	-0.51	0.30	0.01
Japan	1.56	0.47	0.47	0.06	-0.47	0.24	0.73	0.24	0.26	0.24	0.05
Macao-China	-0.07	0.10	0.47	0.07	-0.04	0.08	-0.03	0.04	-0.07	0.08	0.02
Malaysia	-0.10	0.31	0.41	0.06	0.29	0.32	-0.04	0.13	0.25	0.29	0.00
Montenegro	0.05	0.22	0.46	0.07	0.08	0.14	0.02	0.10	0.10	0.12	0.01
Netherlands	0.07	0.21	0.27	0.10	-0.60	0.14	0.02	0.05	-0.58	0.16	0.01
Norway	0.23	0.20	0.34	0.11	-0.52	0.18	0.08	0.06	-0.44	0.16	0.03
Poland	0.05	1.43	0.46	0.12	0.82	0.57	0.02	0.66	0.84	0.30	NA
Portugal	0.23	0.21	0.57	0.08	-0.20	0.20	0.13	0.12	-0.07	0.18	0.04
Russian Federation	0.01	0.19	0.26	0.08	-0.16	0.14	0.00	0.05	-0.15	0.15	0.00
Serbia	-0.06	0.14	0.43	0.06	-0.10	0.15	-0.02	0.06	-0.13	0.14	0.01
Shanghai-China	-0.99	0.59	0.40	0.10	-0.51	0.35	-0.40	0.24	-0.91	0.46	0.04
Singapore	-0.15	0.14	0.39	0.13	0.15	0.12	-0.06	0.07	0.08	0.11	0.03
Slovak	-0.65	0.23	0.33	0.08	0.49	0.34	-0.21	0.08	0.28	0.30	0.02

Republic											
Slovenia	-0.32	0.21	0.79	0.12	0.00	0.24	-0.25	0.18	-0.25	0.18	0.10
Spain	-0.26	0.20	0.54	0.10	-0.10	0.21	-0.14	0.12	-0.24	0.23	0.05
Sweden	0.14	0.22	0.43	0.16	-0.26	0.15	0.06	0.11	-0.20	0.14	0.03
Switzerland	0.27	0.32	0.51	0.07	-0.51	0.32	0.14	0.16	-0.37	0.28	0.02
Turkey	0.08	0.39	0.34	0.07	-0.05	0.36	0.03	0.13	-0.02	0.45	0.00
United Arab Emirates	0.54	0.10	0.53	0.06	0.06	0.08	0.29	0.06	0.34	0.07	0.15
United Kingdom	-0.01	0.16	0.51	0.08	-0.44	0.17	0.00	0.08	-0.44	0.16	0.00
United States of America	0.36	0.14	0.31	0.08	-0.32	0.11	0.11	0.05	-0.21	0.10	0.05
Uruguay	0.66	0.40	0.31	0.07	0.16	0.32	0.20	0.13	0.36	0.33	0.02

Note. a: effect of migration on exploration. b: effect of exploration on CPS. NA indicates that κ^2 could not be calculated due to a zero covariance between migration background and exploration.

ANHANG C: ARBEIT 3

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020). Exploring Behavioral Patterns during Complex Problem Solving. *Journal of Computer Assisted Learning*. Advance online publication. https://doi.org/10.1111/jcal.12451

Received: 14 November 2019 | Revised: 31 March 2020 | Accepted: 20 April 2020

DOI: 10.1111/jcal.12451

ARTICLE

Journal of Computer Assisted Learning WILEY

Exploring behavioural patterns during complex problem-solving

Beate Eichmann[1] | Samuel Greiff[2] | Johannes Naumann[3] | Liene Brandhuber[4] | Frank Goldhammer[1]

[1]Centre for International Student Assessment (ZIB), Educational Quality and Evaluation, DIPF Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany

[2]Cognitive Science and Assessment, Department of Behavioral and Cognitive Sciences, University of Luxembourg, Esch-sur-Alzette, Luxembourg

[3]School of Education, Institute of Educational Research, University of Wuppertal, Wuppertal, Germany

[4]Institut für Allgemeine Erziehungswissenschaften, Goethe University Frankfurt, Frankfurt am Main, Germany

Correspondence
Beate Eichmann, DIPF | Leibniz Institute for Research and Information in Education, Rostocker Straße 6, 60323 Frankfurt am Main, Germany.
Email: beate.eichmann@dipf.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Numbers: 01LSA1504A, 01LSA1504B; Fonds National de la Recherche Luxembourg, Grant/Award Number: The Training of Complex Problem Solving; "TRIOPS"

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1111/jcal.12451.

Abstract

In this explorative study, we investigate how sequences of behaviour are related to success or failure in complex problem-solving (CPS). To this end, we analysed log data from two different tasks of the problem-solving assessment of the Programme for International Student Assessment 2012 study (n = 30,098 students). We first coded every interaction of students as (initial or repeated) exploration, (initial or repeated) goal-directed behaviour, or resetting the task. We then split the data according to task successes and failures. We used full-path sequence analysis to identify groups of students with similar behavioural patterns in the respective tasks. Double-checking and minimalistic behaviour was associated with success in CPS, while guessing and exploring task-irrelevant content was associated with failure. Our findings held for both tasks investigated, from two different CPS measurement frameworks. We thus gained detailed insight into the behavioural processes that are related to success and failure in CPS.

KEYWORDS

complex problem-solving, exploration, log data, PISA, sequence analysis

1 | INTRODUCTION

Our world and our society are becoming increasingly complex. In particular the fast development of technology confronts people more and more frequently with challenges in dealing with unknown situations (e.g., installing a smart TV, using driving assistance technology, using a new app on the smartphone) (Autor, Levy, & Murnane, 2003). Also in non-technical contexts complexity increases. Globalization connects people all around the world, leading to global markets and organizations in which interests of more interdependent parties have to be managed than in small, local structures (Wilpert, 2009). The ability to cope with novel situations of these kinds is addressed in

research on complex problem-solving (CPS). One common definition of CPS that we use in this article is put forward by Frensch and Funke (1995, p. 36):

> CPS occurs to overcome barriers between a given state and a desired goal state by means of behavioural and/or cognitive, multistep activities. The given state, goal state, and barriers between given state and goal state are complex, change dynamically during problem-solving, and are intransparent. The exact properties of the given state, goal state, and barriers are unknown to the solver at the outset. CPS implies the efficient interaction between a solver and the situational requirements of the task, and involves a solver's cognitive, emotional, personal, and social abilities and knowledge.

The skill to solve complex problems can be regarded as a so-called 21st century skill—a skill becoming increasingly relevant for both work and private life in the 21st century (Binkley et al., 2012). Therefore, the question arises which "multistep activities" (Frensch & Funke, 1995, p. 36) and behavioural patterns, respectively, lead to success or failure in CPS. Gaining knowledge about crucial processes in CPS is fundamental to making students better problem solvers and prepare them for the challenges of the future. Previous studies investigated the effects of different behaviours on success in CPS (Eichmann, Goldhammer, Greiff, Pucite, & Naumann, 2019; Greiff, Niepel, Scherer, & Martin, 2016; Schult, Stadler, Becker, Greiff, & Sparfeldt, 2017; Sonnleitner, Brunner, Keller, & Martin, 2014). However, these studies mostly used single unit measures of CPS behaviour, for example, they investigated how the occurrence or frequency of certain behaviours or the time spent on (parts of) a task is related to success. Extending these lines of research, in the present study we want to get more comprehensive insights into the relation between behavioural characteristics and success in CPS by considering complete sequences of certain behaviours. Thereby we want to identify effects of patterns or compositions of behaviours. To achieve this, we will use full-path sequence analysis to identify and group together similar behavioural sequences. According to the definition by Frensch and Funke (1995) mentioned above, multistep interactions (i.e., sequences of behaviour) are key to CPS.

1.1 | Assessment of CPS behaviour

CPS research regularly employs computer simulations of real world problems (e.g., handling an MP3 player, OECD, 2013) to assess CPS skills. Since in computer-based assessment the interactions of participants with the assessment system are recorded, this behavioural data can then be analysed, and inferences about relations between behaviour and success in CPS can be made. There are two widely used CPS assessment frameworks that are explained in detail in the method section of this article (Funke, 2001; Greiff, Fischer, Stadler, & Wüstenberg, 2001). The advantage of these formal frameworks is that they make results from different studies comparable and allow for a systematic description of the underlying task structures (Greiff, Wüstenberg, & Funke, 2012). CPS tasks built under both frameworks are usually highly interactive and lead to rich log data. In log data the behaviour of the problem solver while working on a task is stored. Therefore, this data enables the investigation of behaviour while solving complex problems. In the following sections, we will describe and discuss the different approaches and results of previous research on CPS log data.

1.2 | Top-down approaches to investigate CPS log data

As mentioned before, previous research often used single unit measures to investigate the relation between behaviour and success in CPS (Kroehne & Goldhammer, 2018; Naumann, Goldhammer, Rölke, & Stelter, 2014; Richter, Naumann, & Noller, 2003). In this top-down approach, theory-driven hypotheses about the relations between certain behaviours and success in CPS are formulated. These behavioural states are then identified by events that are included in the log data (Kroehne & Goldhammer, 2018). Single unit measures rely on single events and do not take into account sequential information (Richter et al., 2003). They are derived, for instance, by determining the (cumulated) time in a certain state or the frequency of a certain event or type of event. Naumann et al. (2014) investigated the relation between the number of interactions and success in technology-based problem-solving. They assumed that in everyday technology-based problems most people not solving the problems behave too passive (rather than too active). Their results revealed that low achieving students indeed often show too little interaction with the problem at hand. Moreover, several studies showed a positive relation between the amount of exploration and success in CPS (Dormann & Frese, 1994; Eichmann, Goldhammer, Greiff, Brandhuber, & Naumann, 2018). Exploration can serve several purposes in CPS. First, exploration can be required to gather necessary information to solve a problem. Second, exploration can be non-targeted; that is exploration of task-irrelevant information. For example, if the problem requires the problem solver to buy a subway ticket on a ticket machine, interacting with buttons for bus tickets would be regarded as non-targeted exploration, since these interactions are not directly goal-related. However, non-targeted exploration can serve the purpose of getting to know the problem space and can therefore support the problem solver to build a mental model of the problem (Dormann & Frese, 1994). Bell and Kozlowski (2008) found a positive relation between exploration and metacognitive activity. They argue that metacognitive activity also facilitates CPS. However, Bell and Kozlowski (2008) did not differentiate between exploration of necessary information task-irrelevant information. Greiff et al. (2016) found a low intervention frequency to be advantageous in CPS. Hence, they argue that CPS benefits from planned behaviour. Accordingly, Eichmann et al. (2019) showed that, especially in the beginning of a CPS process, taking time to plan ahead has a positive impact on

success. They also argued that in the course of CPS, time allocation into phases of higher and lower activity plays an important role, especially for difficult problems. A quite well investigated CPS strategy applicable to certain CPS tasks is the vary-one-thing-at-a-time strategy (VOTAT). The VOTAT strategy implies manipulating single variables, while all other possible input variables are kept constant. Thus, the influence of the manipulated variable on other variables can be investigated and knowledge about the problem can be generated, which has a positive effect on success in CPS (Greiff, Wüstenberg, & Avvisati, 2015; Tóth, Rölke, Greiff, & Wüstenberg, 2014; Wüstenberg, Greiff, Molnár, & Funke, 2014). The VOTAT strategy has also received much attention in research on science inquiry (Apedoe & Schunn, 2013; Jirout & Zimmerman, 2015). Problem solvers' use of the VOTAT strategy is usually also operationalized by count indicators that reflect whether and to what extent the strategy was used.

The top-down approach of using single unit measures to analyse log data has the advantage that theory-based assumptions about effects of behaviour can be investigated. Relations between single unit measures (frequencies or durations of single behaviours) and success in CPS can easily be investigated using the approach adopted in many of the studies mentioned before. However, this approach reduces behavioural sequences to single numbers. Through this reduction of data effects of combinations or sequences of behaviours (i.e., the exact order of actions) might be overlooked and important information might get lost. Of course, information reduction can also be useful to reduce noise in the data. However, Richter et al. (2003) argue that single unit measures might lead to similar measures for in fact very different behaviours since they do not take into account sequential information. Therefore, they recommend the (additional) use of sequential measures. The differences between the aforementioned top-down approaches and bottom-up approaches used with sequential measures are depicted in Figure 1. In the top-down approach, theory-based indicators are formed (e.g., single unit measures), in the bottom-up approach regularities in the data (e.g., clusters of behavioural sequences) are interpreted on the basis of theoretical assumptions. Bottom-up approaches used to analyse log data will be explained in more detail in the following section.

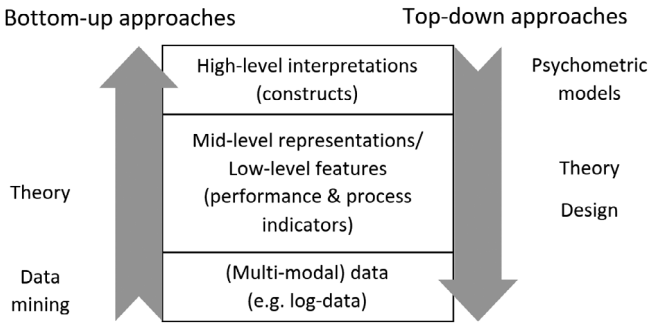

FIGURE 1 Comparison of top-down and bottom-up approaches to identify evidence for drawing inferences about a construct. *Source:* Adapted from Mislevy (2019, p. 35)

1.3 | Bottom-up approaches to investigate CPS log data

To overcome the limitations of single unit measures, in the recent literature methods to investigate (sub)sequences of behaviour were applied (He & von Davier, 2015; Stadler, Fischer, & Greiff, 2019). He and von Davier (2015) used the *n*-gram approach to identify patterns of behaviour related to success in CPS. The *n*-gram approach decomposes CPS behaviour into small subsequences and analyses the relation between the frequency of these subsequences and success in CPS. In this data-driven approach log data is analysed without prior hypotheses about specific behaviours. The substantive interpretation of the results will then take place a posteriori. He and von Davier (2015) found actions that were *not* part of the shortest path to success to be associated with *not* solving a complex problem, while more goal-directed actions were associated with solving it. They investigated all the possible actions in the investigated task as distinct behaviours obtaining 27 different behaviours, which were combined into 144 bigrams and 257 trigrams. Using the same approach Stadler et al. (2019) showed behaviour sequences were more likely to be related to success if they could be assumed to generate less cognitive load. They argue that cognitive load is lowest if the problem solver follows the direct path to the correct solution. In contrast to the study of He and von Davier (2015), Stadler et al. (2019) pre-coded the actions of their test takers according to two distinct behaviour categories (manipulating variables and annotating the observed results in a CPS task). They argue that the direct path to success (generating the least cognitive load) is characterized by annotating results immediately after every variable manipulation instead of performing several manipulations subsequently. Thus, still using a bottom-up approach they integrated the top-down element of using a (rather task-driven) coding for their data. Both studies showed goal-directed behaviour to be beneficial in CPS, while non-targeted exploration was found to be detrimental. However, this finding runs contrary to the results of Dormann and Frese (1994) and Eichmann et al. (2018) who reported a positive relation between exploration (both goal-directed and non-targeted) and success in CPS. Therefore, the question remains under which circumstances the respective effects arose.

The apparent ambiguity in these findings concerning the usefulness of exploration may be a result of the varying definitions of exploration in contrast to goal-directed behaviour. He and von Davier (2015) and Stadler et al. (2019) emphasize the shortest path to the successful solution to be beneficial. They did not consider repeated actions in this regard and therefore concluded parsimonious behaviour to be beneficial. Eichmann et al. (2018) also considered the shortest path to success as goal-directed. However, all interactions that went beyond this shortest path including repeated recapitulation (of goal-directed parts of the task) were regarded as exploration. Thus, the effects of goal-directed exploration (i.e., repeated goal-directed interactions) and non-targeted exploration could not be disentangled. However, investigating repeated goal-directed interactions and non-targeted exploration separately might be a way to clarify the usefulness of these behaviours in CPS. As Greiff, Molnár, Martin,

Zimmermann, and Csapó (2018) argue, not only the quantity but also the quality of exploration might be related to CPS performance. Moreover, using the full path of behaviour might help to clarify the usefulness of not only single instances of these behaviours but also of behaviour patterns that might be more complex than three or four subsequent actions, which is the maximum length of *n*-grams used by He and von Davier (2015) and Stadler et al. (2019).

The ambiguous results on the effects of exploration and goal-directed behaviour between studies using the *n*-gram approach and studies using single unit measures demand a method to clarify the role of exploration and goal-directed behaviour in CPS. By combining the advantages of theory-driven top-down approaches and data-driven bottom-up approaches we aim at taking an even closer look at behavioural processes in CPS.

1.4 | A top-down bottom-up mixed approach to investigate CPS log data

To take this closer look, in this study we apply full-path sequence analysis (Gabadinho, Ritschard, Müller, & Studer, 2011), an exploratory approach that does not only take into account sub-sequences of CPS but the whole behavioural sequence of every problem solver. In full-path sequence analysis, complete sequences (of behaviour) are clustered according to their similarity leading to clusters of similar CPS behaviours. With these clusters we hope to identify possibly heterogeneous behaviours, which alike might lead to either correct or false solutions in CPS. The aforementioned sequence analysis methods were originally used for comparing DNA sequences and use string matching algorithms to determine the similarity of sequences (Abbott & Forrest, 1986). There are different string matching algorithms available that take into account different attributes of the sequences to be compared. Comparable to the *n*-gram approach, sequence analysis methods can be applied to rather raw (see He & von Davier, 2015) or pre-coded (see Stadler et al., 2019) log data.

In this study, we use behavioural categories that have been shown to be relevant to success in CPS in previous research for coding our data. Therefore, we integrate the top-down approach of theory-driven single unit measures (through pre-coding) with the bottom-up approach of exploratory full-path sequence analysis. We distinguish non-targeted exploration behaviour, which has been shown to be positively related to success in CPS by Dormann and Frese (1994) and Eichmann et al. (2018), from goal-directed behaviour (including goal-directed exploration), which was found to be most positively related to success in the studies of He and von Davier (2015) and Stadler et al. (2019). According to Dormann and Frese (1994) exploration denotes metacognitive activities and helps building a mental model of the problem at hand. In contrast, goal-directed behaviour (as we defined it) reflects an efficient processing of the tasks content. Since from previous research it is unclear whether parsimony is beneficial for CPS performance (Eichmann et al., 2018; He & von Davier, 2015; Stadler et al., 2019), we also distinguish between initial and repeated actions. As Wirth (2004) argues, repeating actions could be an attempt to integrate information that has been identified before. Repeating goal-directed actions could therefore reflect thoroughness, while repeating non-targeted exploration could reflect an overestimation of the relevance of the inspected information. Students' behaviour in a specific CPS task could therefore indicate general student characteristics such as perseverance or motivational states. Therefore, students' behaviour might not only predict success in this very task but might also be an expression of this student's overall CPS performance. Therefore, we want to investigate the relation between students' behaviour and both their performance in the very item, in which the behaviour was shown, and their overall CPS performance. Sequence analysis methods have the advantage that they take into account both the frequency of behaviours as well as the order of behaviours throughout the whole behavioural path (Gabadinho et al., 2011; Studer, Ritschard, Gabadinho, & Müller, 2011). Through this we hope to clarify the circumstances under which parsimony, non-targeted exploration, and goal-directed behaviour are beneficial for successful CPS.

1.5 | Hypotheses and research questions

Although Dormann and Frese (1994) and Eichmann et al. (2018) found positive effects of non-targeted exploration, we expect these effects to be possibly confounded with repeated goal-directed behaviour. Since He and von Davier (2015) argue that non-goal-directed behaviour should be detrimental, we expect the positive effect found by Dormann and Frese (1994) and Eichmann et al. (2018) to be due to repeated goal-directed actions being part of their measure of exploration. Therefore, our hypotheses are as follows:

Hypothesis 1 Non-targeted exploration (both initial and repeated) is more frequent among false CPS solutions than correct CPS solutions.

Hypothesis 2 Goal-directed behaviour (both initial and repeated) is more frequent among correct CPS solutions than false CPS solutions.

In addition, we want to explore the complex behaviour patterns that result in success or failure in CPS. Thus, we formulated the following research questions:

Research question 1 Which clusters of behavioural patterns (in terms of complete CPS behaviour sequences) are related to success or failure in CPS? To the best of our knowledge, previous research did not address the relation between complete CPS processing paths and success in the respective CPS tasks.

Research question 2 Can clusters of behavioural patterns and their relation to success in CPS be generalized across different task frameworks? The question of generalizability of these relations between behaviour sequences and success in CPS across different CPS task frameworks has not been addressed by previous research.

2 | METHOD

2.1 | Sample

We used data from the computer-based assessment of the Programme for International Student Assessment (PISA) 2012. In the PISA study, the competencies of 15-year-olds are assessed in several countries. We used the data of those students who worked on at least one of the two tasks for our analysis. The sample consisted of N = 30,098 students from 42 countries; 50.37% were female.

2.2 | Instruments

There are two widely used frameworks to measure CPS skills: Finite State Automata (FSA) and Linear Structural Equations (LSE) (Funke, 2001; Greiff, Fischer, Stadler, & Wüstenberg, 2014). FSA are characterized by a finite number of distinct states the system can attain. The problem solver can use a defined set of operators to switch between these states. An example for such a system is a ticket machine on which the problem solver can use different buttons (=operators) to navigate through several options (e.g., daily ticket or individual trips) for buying tickets (=states).

Tasks from the LSE framework are based on a number of interrelated input (exogenous) and output (endogenous) variables. The relations between the variables are unknown to the problem solver. By manipulating the exogenous and observing the endogenous variables the relations between them can be investigated. An example for such a system is the control of room temperature and humidity (=endogenous variables) using the sliders of a climate control (=exogenous variables). These two frameworks can be used to design CPS tasks covering a wide range of real world problems within different fields of knowledge. Also, these frameworks allow for intentional, theory-driven manipulation of item difficulty (Stadler, Niepel, & Greiff, 2016).

To cover both the LSE and the FSA framework we chose one CPS task from either framework for analysis. Both tasks were released by the OECD. We used two tasks that provide prototypical instances for LSE and FSA type problems, respectively, and that had a comparable

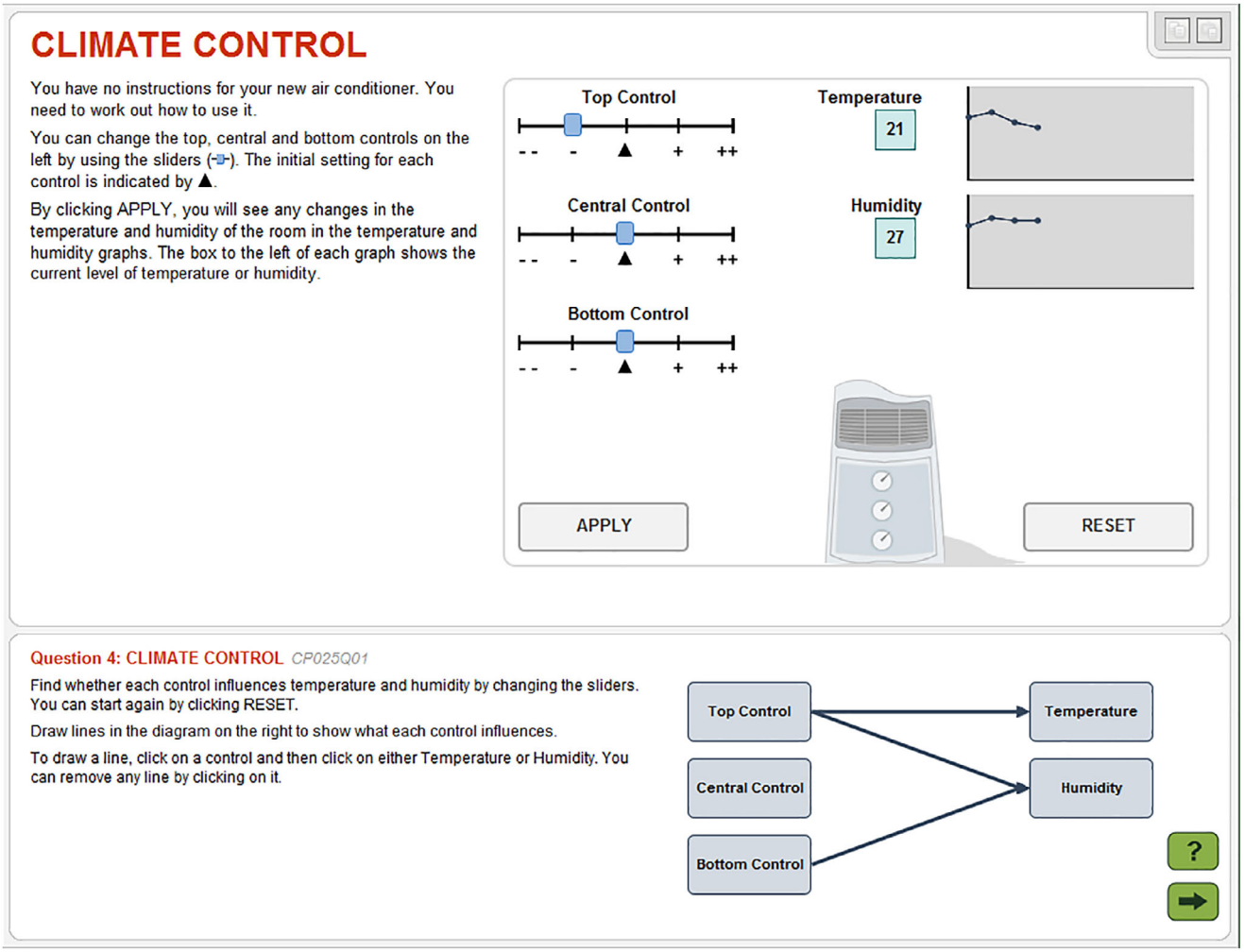

FIGURE 2 The climate control task from PISA 2012 after the arrows in the diagram have been drawn. PISA, Programme for International Student Assessment [Colour figure can be viewed at wileyonlinelibrary.com]

number of minimum actions required to solve the task. In both tasks it was straightforward to distinguish between goal-directed behaviour and non-targeted exploration. From the LSE framework we chose the climate control task (see Figure 2). In this task, students were required to investigate the relations between exogenous and endogenous variables and then visualize these relations by drawing lines in a diagram. The exogenous variables were control sliders regulating the endogenous variables, which were temperature and humidity. The goal of this task was to obtain a diagram that correctly represents all existing relations between exogenous and endogenous variables. To obtain a correct solution in this task a minimum of six goal-directed actions was required. From the FSA framework, we chose the tickets task (see Figure 3). In this task, students were required to navigate through the states of a ticket machine to reach a desired goal state. The goal was to buy the cheapest ticket available considering particular requirements. Students had to compare and revisit several states to decide which ticket was the cheapest. To compare all relevant and find the correct ticket a minimum of seven goal-directed actions and one reset was required. We used the students' responses on all other 25 CPS tasks from the PISA 2012 assessment to determine their overall CPS skills. We did not use the plausible values for problem-solving from the PISA 2012 database for two reasons: First, the plausible values include both complex and analytical problem-solving performance. However, we only want to investigate complex problem-solving performance. Second, the plausible values also include the raw scores of the two items analysed in our study. Therefore, investigating the relation between performance in our two items and PISA's plausible values would lead to an overestimation of the relation between behaviour and performance. For details about the PISA problem-solving assessment see OECD (2013).

2.3 | Procedure

Problem-solving was part of the optional computer-based assessment in PISA 2012. In the participating countries, the computer-based assessment was carried out after the paper-based assessment. Students first received a tutorial to practice the required actions with the computer-based assessment environment to eliminate any effects of

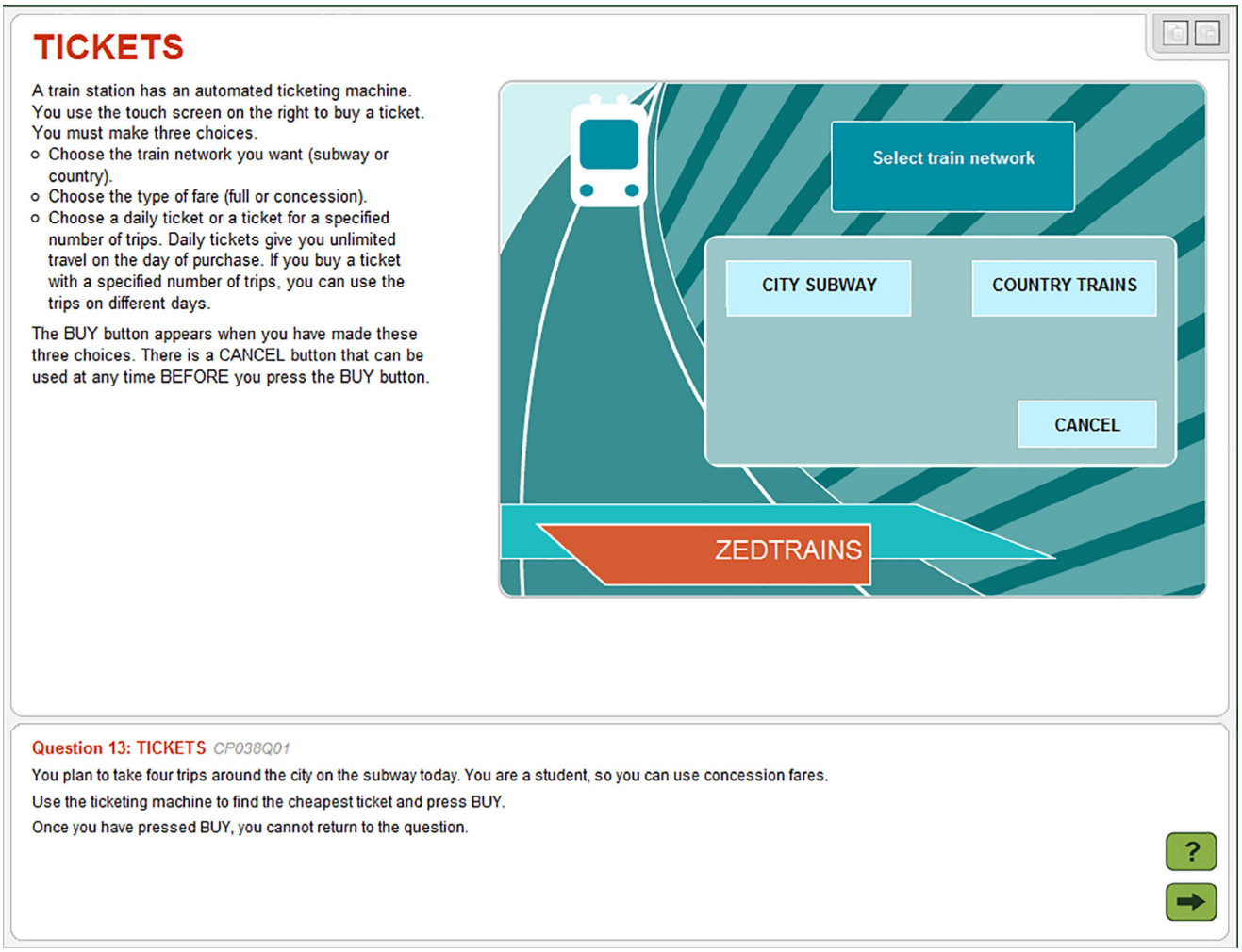

FIGURE 3 The tickets task from PISA 2012. PISA, Programme for International Student Assessment [Colour figure can be viewed at wileyonlinelibrary.com]

students' ICT skills. According to the booklet design of PISA, students received either two problem-solving clusters or one problem-solving cluster and one task cluster from a different assessment domain. Students were then given 20 min time to complete each computer-based cluster. Each problem-solving cluster contained four problem-solving units. The problem-solving units consisted of two to three tasks each (OECD, 2014a). The two tasks we used for our analyses were located on position 2 (tickets task) and position 4 (climate control task) of their cluster. Depending on whether this cluster was administered as first or second cluster, the tasks were either in an early position of the test or in a middle position.

2.4 | Data preparation

Before coding the log data, we deleted log events that were not caused by student action (e.g., log events that mark the loading or unloading of a task). As mentioned above, we coded all remaining log events in our data separately as categories of behaviour. As discussed earlier, previous research points out the importance of exploration for success in CPS (Bell & Kozlowski, 2008; Dormann & Frese, 1994; Eichmann et al., 2018) but also the positive effects of parsimonious, goal-directed behaviour (He & von Davier, 2015; Stadler et al., 2019). Therefore, we chose to use the categories non-targeted exploration and goal-directed behaviour for our sequence analysis. We defined goal-directed behaviour as every interaction *necessary* for the students to solve the respective task correctly (i.e., every interaction that is part of the shortest path to task success given the knowledge the problem solver has at the beginning of each task). Therefore, goal-directed behaviour also includes exploration that is required to solve the task. In contrast to this, non-targeted exploration was operationalized as every interaction *not* necessary to solve the task. The distinction between non-targeted exploration and goal-directed behaviour was implemented differently for the tasks from the LSE and FSA framework. In FSA tasks it is possible to simply distinguish between states of the problem that are required to be visited for a correct solution and states that were not. In LSE tasks this is not the case. Therefore, we decided to define the use of the VOTAT strategy, the drawing of a correct line, and the deletion of a wrong line in the diagram as goal-directed behaviour in the climate control task (Wüstenberg, Greiff, & Funke, 2012). Other interactions in the climate control task (e.g., manipulating multiple variables at a time, drawing a wrong or deleting a correct line in the diagram) were coded as non-targeted exploration. We further refined the two categories goal-directed behaviour and non-targeted exploration by also distinguishing whether an interaction was performed for the first time or repeatedly. Both tasks contained a reset button, which would restore the initial state of the task. Pressing the reset button could not be categorized as goal-directed or non-targeted exploration, since resetting could be both part of non-targeted exploration or part of goal-directed interaction. Therefore, we defined resetting as a unique category. Thus, we ended up with five categories of CPS behaviour: initial goal-directed behaviour, repeated goal-directed behaviour, initial non-targeted exploration, repeated non-targeted exploration and resetting.

2.5 | Data analysis

We divided the dataset by task (climate control vs. tickets) and by the correctness of the given response (correct vs. false). By dividing the data into correct and false trials, we wanted to obtain behaviours that can be clearly assigned to correct or false responses, rather than obtaining more or less successful behaviours. In the climate control task 51.08% of the students gave a correct response. In the tickets task, 43.04% of the students did so. The following analyses were performed for the four resulting subsets of the data separately.

For Hypotheses 1 and 2 we conducted chi-squared tests to compare the relative frequencies of (initial and repeated) non-targeted exploration and (initial and repeated) goal-directed behaviour between correct and false responses in the two tasks, respectively. For the sequence analysis, we determined the differences between the sequences of behaviour categories of the students using optimal matching and the R package TraMineR (R Core Team, 2016; Studer & Ritschard, 2016). The optimal matching algorithm determines the dissimilarity between sequences by calculating the costs of transferring one sequence into the other. There are two types of costs to be specified. The costs for inserting or deleting an element of the sequence (indels), which reflect differences in sequence length, and the costs fur substituting one sequence element with another element. We chose indels = 2 and the substitution cost matrix shown in Table 1 for our analysis. The substitution costs reflect the theoretical similarity between the behaviour categories (e.g., initial and repeated non-targeted exploration are more similar to each other than initial non-targeted exploration and initial goal-directed behaviour). Our indels equal the maximum of the substitution costs, so a difference in length between two sequences would result in the same difference value as a difference between one behaviour category and a very dissimilar one. Therefore, both sequence lengths as well as qualitative differences between sequences are taken into account to determine the dissimilarity between sequences. We also normalized the dissimilarity between the sequences dividing it by the length of the longer of each two sequences to account for potentially larger (non-normalized) dissimilarity between longer sequences (Gabadinho et al., 2011). Note that the comparison between sequences refers to the order of interactions and not to timing (i.e., two sequences are regarded as being identical if they contain the same interactions in the same order, no matter if the interactions were performed with different speed).

Based on the differences of students' sequences, we conducted a hierarchical cluster analysis using the Ward algorithm (Studer, 2013). We used the PISA 2012 final student weights in the analysis to account for oversampling. We used the normalized point-biserial correlation (PBC), average silhouette width (ASW) and Hubert's C index (HC) as quality criteria to determine the optimal number of clusters. The PBC measures the capacity of a clustering solution to reproduce the differences between sequences obtained through string matching. The ASW compares the average weighted distance of a cluster member from other members of the same cluster with its average weighted distance from the closest other cluster. The HC reflects the difference between the obtained cluster solution and the best cluster

TABLE 1 Substitution cost matrix for optimal matching

	Initial exploration	Repeated exploration	Initial goal-directed behaviour	Repeated goal-directed behaviour	Resetting
Initial exploration	0	1	1.5	2	1.5
Repeated exploration	1	0	2	1.5	1.5
Initial goal-directed behaviour	1.5	2	0	1	1.5
Repeated goal-directed behaviour	2	1.5	1	0	1.5
Resetting	1.5	1.5	1.5	1.5	0

solution that could have been obtained with the given dataset and number of clusters. While PBC and ASW should be maximized to obtain the optimal clustering solution, HC should be minimized to do so (Studer, 2013). PBC, ASW and HC take into account different properties of the cluster solutions. Considering these different indices we aim at a well-balanced evaluation of the different cluster solutions. For our four different data subsets we tested hierarchical clustering with two to eight clusters, respectively and chose the optimal solutions according to our quality criteria.

To compare the obtained clusters with respect to students' overall CPS skills, we used the responses of the students on the other 25 CPS items of the PISA 2012 assessment. The responses were coded as no credit, partial credit, full credit or not reached. We recoded not reached items as no credit and fitted a one-parameter logistic (1PL) partial credit item response theory (IRT) model to the response data to obtain weighted likelihood estimators (WLEs) of students' overall CPS skills using marginal maximum likelihood estimation of the TAM package (Robitzsch, Kiefer, & Wu, 2019). Maximum likelihood estimation allows to compute unbiased means of ability estimates (even though the variance might be overestimated) (Mislevy, Beaton, Kaplan, & Sheehan, 1992). Since we used WLEs based on maximum likelihood estimation to compare group means only, we refrained from the more complex analysis approach using plausible values. Due to PISA's rotated block design, there were missing responses by design in all CPS items. The WLE scale is centred so its mean is zero. Therefore, negative WLEs represent CPS skills below average. We used the PISA final student weights to account for the stratified sampling (OECD, 2014b). Subsequently, we applied analysis of variance to compare the mean CPS skills across clusters. To obtain group-wise comparisons, we used Tukey Honest Significance Difference test, which controls for Type I error inflation (Field, Miles, & Field, 2013). We also compared the clusters regarding their occurrence depending on tasks' positions in the test (early vs. middle position). Therefore, we used chi-squared tests to compare if clusters occur significantly more often at an early or a middle position in the test. For all chi-squared tests we calculated the effect size φ using the DescTools package in R (Signorell, Andri et mult. al., 2019). According to the conventions of Cohen (1988), a φ value of 0.1 is considered a small effect, 0.2 a medium effect and 0.3 a large effect. For all Tukey tests we calculated Cohen's d using the psych package in R (Revelle, 2018). According to the conventions of Cohen (1988), a d value of 0.2 is considered a small effect, 0.5 a medium effect and 0.8 a large effect.

3 | RESULTS

The results of the chi-squared tests comparing the relative frequencies of goal-directed behaviour and non-targeted exploration are shown in Tables 2 and 3. In line with Hypothesis 1, both initial and repeated non-targeted exploration was more frequent among false responses. The results for goal-directed behaviour were only significant for repeated goal-directed behaviour in the tickets task (Table 3). This indicates that in the tickets task repeated goal-directed behaviour was more frequent among correct responses. Therefore, Hypothesis 2 is only supported in this particular case.

The quality criteria PBC, ASW and HC according to the different numbers of clusters are shown in Figure 4. Since not all the quality criteria favoured the same solution in all cases, we decided on those solutions that were favoured by at least one quality criterion while also showing good results for the other two criteria.

Following this rule, we decided on the 5-cluster solution for false solutions in the climate control task, which reflects the maximum of the PBC, a local minimum of HC and a medium value for ASW. A high value of PBC indicates that the partition reflects the patterns of dissimilarities between sequences quite well. A low value of HC reflects a favourable ratio of within- and between-cluster dissimilarities (Studer, 2013). We chose a 3-cluster solution for correct solutions in the climate control task, again maximizing PBC, choosing a local minimum for HC and a medium value for ASW. We chose a 7-cluster solution for false solutions in the tickets task, this time optimizing both PBC and HC while ASW had a value close to its maximum. A high value of ASW reflects a good ratio of sequences' similarities to their cluster members and dissimilarities to members of other clusters. We chose a 6-cluster solution for correct solutions in the tickets task, maximizing PBC and ASW while HC had a value close to its minimum. The resulting clusters are displayed in the following paragraphs.

3.1 | Climate control task

3.1.1 | False solutions

We chose a solution with five clusters of sequences for false solutions in the climate control task. The clusters are depicted in Figure 5. The

TABLE 2 Relative frequencies of different behaviours in the climate control task

	% of behaviour in					
	False responses	Correct responses	χ^2	*df*	*p*	φ
Initial exploration	29.63	14.03	5.57	1	.018	0.36
Repeated exploration	30.78	14.63	5.75	1	.017	0.36
Initial goal-directed behaviour	18.21	31.41	3.50	1	.061	0.27
Repeated goal-directed behaviour	14.17	26.10	3.53	1	.060	0.30

TABLE 3 Relative frequencies of different behaviours in the tickets task

	% of behaviour in					
	False responses	Correct responses	χ^2	*df*	*p*	φ
Initial exploration	24.77	6.78	10.26	1	.001	0.57
Repeated exploration	9.60	0.85	7.32	1	.007	0.84
Initial goal-directed behaviour	40.30	42.36	0.05	1	.821	0.02
Repeated goal-directed behaviour	17.95	37.37	6.82	1	.009	0.35

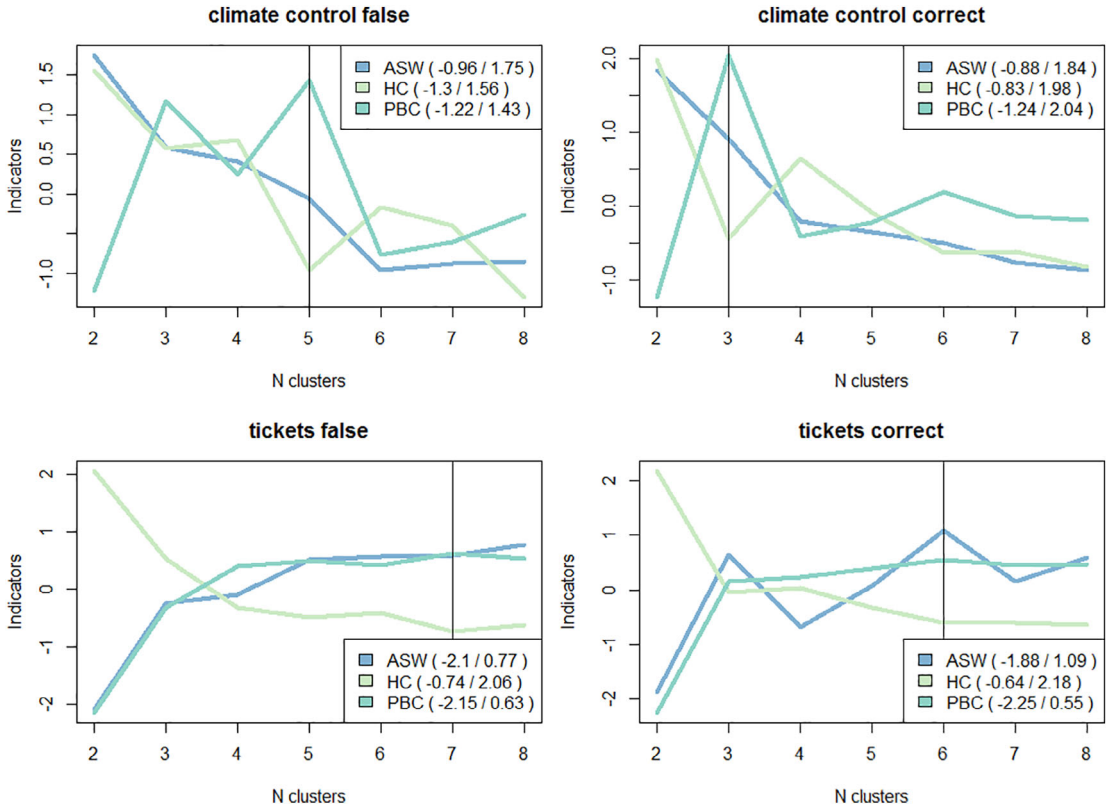

FIGURE 4 *Z*-standardized quality criteria (ASW, PBC and HC) according to number of clusters in the different data sets. Minimum and maximum values are displayed in brackets. The vertical lines mark the chosen solution. ASW, average silhouette width; HC, Hubert's C index; PBC, point-biserial correlation [Colour figure can be viewed at wileyonlinelibrary.com]

figure displays state distribution plots for each cluster that show the relative distribution of behaviour categories at each interaction. Values on the *x*-axis represent the numbered interactions from the behaviour sequences. Values on the *y*-axis represent the relative frequencies of behaviour categories displayed by the students in the respective cluster at the respective interaction. For example, in Figure 5 in the first cluster (top left) roughly 50% of the students showed non-targeted exploration behaviour in their first interaction, about 40% showed goal-directed behaviour in their first interaction, and about 10% reset the task in their first interaction (which has no

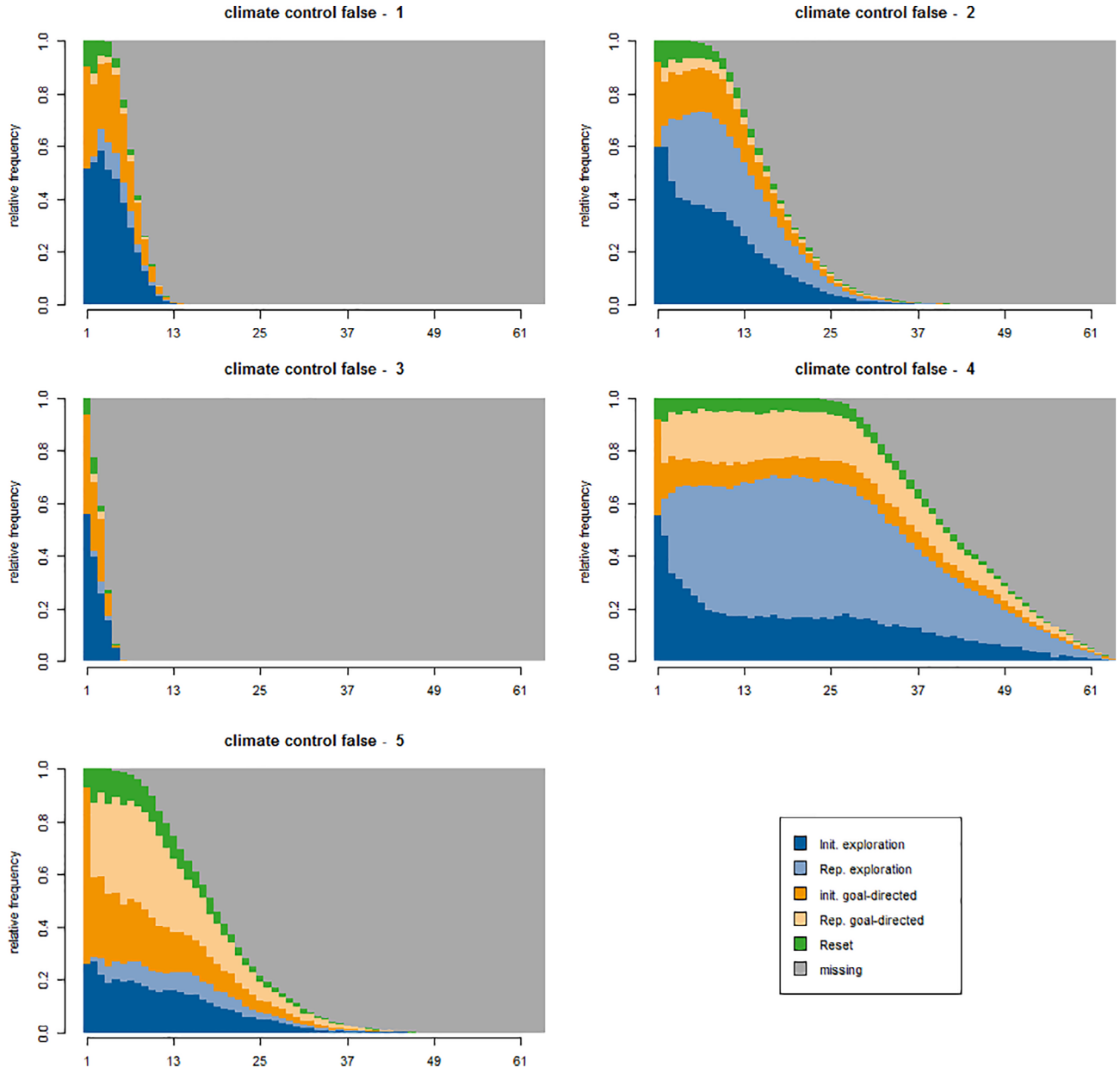

FIGURE 5 State distribution plots for each cluster for false solutions in the climate control task [Colour figure can be viewed at wileyonlinelibrary.com]

effect in the tasks initial state). Missing values in the figure are caused by sequences shorter than the value range of the *x*-axis.

Cluster 1 contains the sequences of 2,672 students (19.31% of all false solutions in the climate control task). They show slightly more non-targeted exploration than goal-directed behaviour. The vast majority of interactions are initial (and not repeated) and the reset button is rarely used. Overall the sequences are quite short with the longest sequence comprising of 15 interactions and an average sequence length of 7.24 interactions. These students seem to unsystematically try out VOTAT and non-VOTAT actions.

Cluster 2 contains the sequences of 4,484 students (32.40% of all false solutions in the climate control task). They show approximately 70% non-targeted exploration behaviour. Especially the non-targeted exploration interactions are often repeated; the reset button is again rarely used. Overall, the sequences are relatively long with the longest sequence containing 49 interactions and an average sequence length of 16.98 interactions. These students also seem to unsystematically try out different actions and engage increasingly in reinvestigation of non-targeted exploratory actions.

Cluster 3 contains the sequences of 1846 students (13.34% of all false solutions in the climate control task). They show approximately 60% non-targeted exploration behaviour. The vast majority of interactions are initial; the reset button is again rarely used. The sequences in this cluster are especially short with the longest sequence consisting of seven interactions and an average sequence length of 2.71

interactions. These students seem to abandon the task quite early resulting in not answering or guessing a false solution.

Cluster 4 contains the sequences of 2,158 students (15.59% of all false solutions in the climate control task). Like cluster 2, they also show approximately 70% non-targeted exploration behaviour. However, in this cluster most interactions (even in the beginning of the sequences) are repeated. The reset button is again rarely used. The sequences in this cluster are the longest sequences among false solutions in the climate control task with an average sequence length of 42.26 interactions. These students reinvestigate the same content again and again. However, most of the investigated content is irrelevant.

Cluster 5 contains the sequences of 2,680 students (19.36% of all false solutions in the climate control task). They show about 60% goal-directed behaviour and about 30% non-targeted exploration. About 10% of the interactions are with the reset button. In this cluster most goal-directed interactions are repeated while most non-targeted exploration is initial. The sequences in this cluster are of similar length as those in cluster 2 with an average sequence length of 18.67 interactions. From the false solutions in the climate control tasks, students in this cluster show the highest proportion of goal-directed behaviour. However, the proportion of initial goal-directed behaviour is quite small. This could indicate an incomplete application of the VOTAT strategy not investigating the effect of *every* input variable in isolation.

The comparison of the clusters regarding overall CPS skills using WLEs shows that all clusters have a negative average estimated skill (see Figure 6). Students guessing or not answering the task (cluster 3) show the lowest average CPS skills, whereas students applying the incomplete approach (cluster 5) show the highest CPS skills. CPS skills increase for clusters with longer sequences and with higher frequencies of goal-directed behaviour. The results of the Tukey Honest Significance Difference test show that significant differences in mean CPS skills exist between all clusters except for the two groups of medium to high sequence length showing mainly non-targeted exploration (clusters 2 and 4) (see Table 4).

The comparison of the clusters regarding their occurrence in early or middle positions in the test reveals that the clusters with very short sequences (clusters 1 and 3) occurred more often in the middle item position (see Table 5). The longer clusters containing mainly exploration (clusters 2 and 4) occur more frequently at the early item position. For cluster 5, no significant difference in occurrence at either position was found.

3.1.2 | Correct solutions

We chose a solution with three clusters of sequences for correct solutions in the climate control task. The clusters are depicted in Figure 7. Cluster 1 contains the sequences of 6,569 students (45.46% of all correct solutions in the climate control task). They show between 10 and 30% non-targeted exploration with a decreasing trend in the course of the problem-solving process and about 70% goal-directed

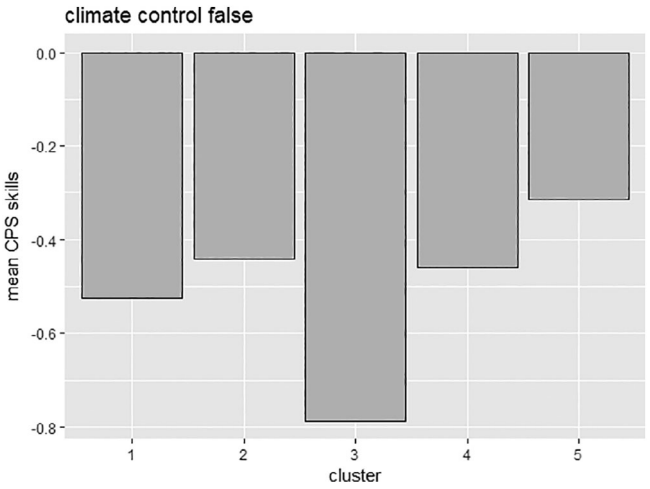

FIGURE 6 Mean CPS skills per cluster for false responses in the climate control task. CPS, complex problem-solving

behaviour. About 20% of the interactions are with the reset button. The majority of interactions are initial. Overall the sequences are of short or medium length with the longest sequence containing 30 interactions and an average sequence length of 13.11 interactions. Since the sequences are quite short and little non-targeted exploration takes place, these students show a quite efficient behaviour.

Cluster 2 contains the sequences of 2,984 students (20.65% of all correct solutions in the climate control task). They show about 30% non-targeted exploration and about 65% goal-directed behaviour. About 5% of the interactions are with the reset button. The majority of interactions is repeated. This cluster contains the longest sequences in the climate control task with sequences up to 80 interactions length and an average sequence length of 48.20 interactions. The large number of repeated goal-directed actions indicates an approach of double-checking relevant information.

Cluster 3 contains the sequences of 4,898 students (33.89% of all correct solutions in the climate control task). They show about 35% non-targeted exploration and about 50% goal-directed behaviour. About 15% of the interactions are with the reset button. The narrow majority of the goal-directed interactions is initial and the narrow majority of the non-targeted exploration is repeated. The sequences in this cluster are rather long with sequences up to 75 interactions length and an average sequence length of 25.69 interactions. These students seem to apply a rather mixed approach with some non-targeted exploration and double-checking but mainly initial goal-directed behaviour.

The comparison of clusters regarding overall CPS skills using WLEs shows that all clusters have a positive average estimated skill (see Figure 8). Students applying the double-checking approach (cluster 2) show the highest average CPS skills. The results of the Tukey Honest Significance Difference test show that the CPS skills of students applying the double-checking approach (cluster 2) are significantly higher than those of students using more efficient (cluster 1) or a mixed approach (cluster 3) (see Table 6). There is no significant difference in overall CPS skill between clusters 1 and 3.

TABLE 4 Comparison of mean CPS ability across clusters for false responses in the climate control task

Compared clusters	Difference	Confidence interval		*p* value	Cohen's *d*
		Lower bound	Upper bound		
2–1	0.08	0.04	0.13	<.001	0.16
3–1	−0.26	−0.32	−0.20	<.001	−0.39
4–1	0.07	0.01	0.12	<.001	0.11
5–1	0.21	0.16	0.26	<.001	0.37
3–2	−0.35	−0.40	−0.29	<.001	−0.53
4–2	−0.02	−0.07	0.03	.869	−0.04
5–2	0.13	0.08	0.17	<.001	0.22
4–3	0.33	0.27	0.39	<.001	0.49
5–3	0.47	0.41	0.53	<.001	0.72
5–4	0.14	0.09	0.20	<.001	0.25

Abbreviation: CPS, complex problem-solving.

TABLE 5 Frequencies of clusters depending on item position for false responses in climate control task

Cluster	Item position in test		χ^2	*df*	*p*	φ
	Early	Middle				
1	1,152	1,349	15.52	1	<.001	0.08
2	2,196	1,859	28.01	1	<.001	0.08
3	500	722	40.33	1	<.001	0.18
4	1,135	855	39.40	1	<.001	0.14
5	1,151	1,246	3.77	1	.052	0.04

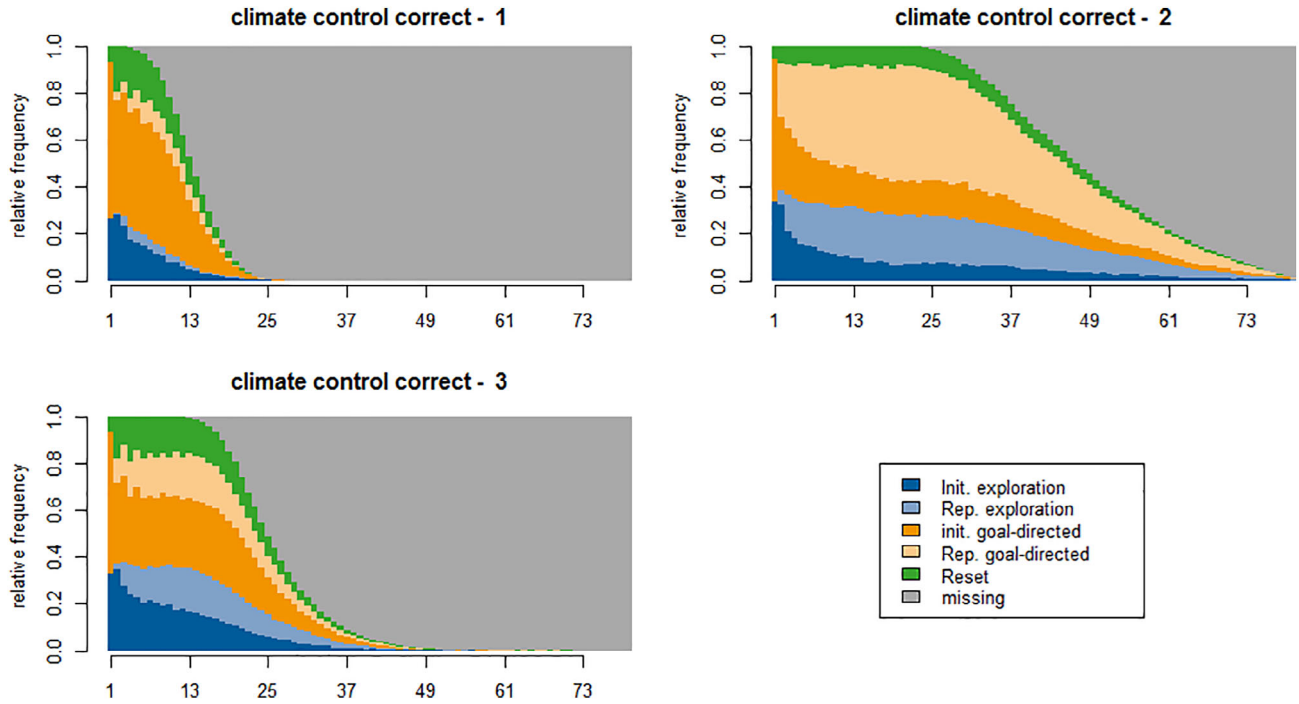

FIGURE 7 State distribution plots for each cluster for correct solutions in the climate control task [Colour figure can be viewed at wileyonlinelibrary.com]

The occurrence of all three clusters differed significantly between the two item positions (see Table 7). While clusters 1 and 2 appeared more frequently in the middle position, cluster 3 was more frequently observed in the early position.

3.2 | Tickets task

3.2.1 | False solutions

We chose a solution with seven clusters of sequences for false solutions in the tickets task. The clusters are depicted in Figure 9. Cluster 1 contains the sequences of 2,334 students (14.61% of all false solutions in the tickets task). They show almost exclusively non-targeted exploration. The reset button was not used by this group. In the beginning all interactions were initial; towards the end almost all interactions were repeated. The sequences are extremely short with the longest sequence containing five interactions and an average sequence length of 4.36 interactions. Since the sequences contain hardly any goal-directed interactions and are also too short to reveal the correct solution, the students seem to apply unsystematic guessing behaviour.

Cluster 2 contains the sequences of 1,668 students (10.44% of all false solutions in the tickets task). They show mainly goal-directed behaviour in their first interaction, afterwards it was almost exclusively non-targeted exploration. This group made little use of the reset button. In the beginning, all interactions were initial; towards the end there were more repeated interactions and resets. The sequences were as short as the sequences in cluster 1 with the longest sequence containing five interactions and an average sequence length of 3.59 interactions. Similar to cluster 1 these students show unsystematic guessing behaviour or do not respond at all.

Cluster 3 contains the sequences of 6,937 students (43.41% of all false solutions in the tickets task). They showed almost exclusively goal-directed behaviour. This group did not use the reset button. During the first three interactions, all interactions were initial; only in the last interaction there were repeated interactions. The sequences are similarly short as the sequences in cluster 1 and 2 with the longest sequence containing four interactions and an average sequence length of 3.98 interactions. Again the short sequences indicate guessing a solution. However, the students seemed to intentionally choose goal-directed actions. Therefore, their behaviour could be called "goal-directed guessing".

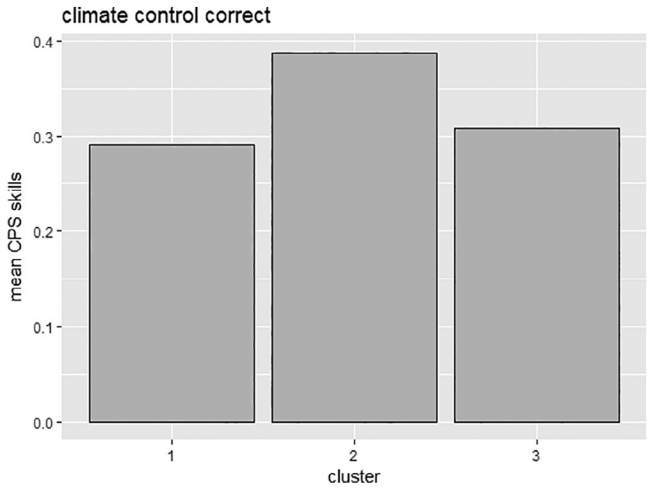

FIGURE 8 Mean CPS skills per cluster for correct responses in the climate control task. CPS, complex problem-solving

Cluster 4 contains the sequences of 1,120 students (7.01% of all false solutions in the tickets task). They showed almost exclusively non-targeted exploration behaviour. This group used the reset button a few times. In the course of the task, the frequency of repeated interactions increases. The sequences are of medium length with the longest sequence containing 18 interactions and an average sequence length of 8.95 interactions. These students seem to be quite persevering in engaging with irrelevant content.

Cluster 5 contains the sequences of 1,078 students (6.75% of all false solutions in the tickets task). They showed almost exclusively goal-directed behaviour. In this group, there is a peak of uses of the reset button at interaction 4. This peak indicates that students navigated to the first ticket option and reset the task to consider further options. Prior to this peak, there were mostly initial interactions. After the peak there were mostly repeated interactions. The sequences are of medium length with the longest sequence containing 12 interactions and an average sequence length of 7.94 interactions. These students show a quite systematic approach. However, instead of comparing different ticket options most students investigated the first ticket twice, which becomes evident in the high proportion of repeated interactions after the peak of resets. Therefore, their approach is rather incomplete.

Cluster 6 contains the sequences of 2,191 students (13.71% of all false solutions in the tickets task). They showed almost exclusively goal-directed behaviour during their first three interactions followed by a peak of reset at interaction 4, similar to cluster 5. However, after the peak there were mostly repeated goal-directed interactions and non-targeted exploration. The sequences are quite long with the longest sequence containing 18 interactions and an average sequence length of 11.66 interactions. These students either compare a relevant ticket with a non-relevant ticket or reinvestigate the first ticket multiple times.

Cluster 7 contains the sequences of 652 students (4.08% of all false solutions in the tickets task). Remarkably, all the sequences in this cluster are identical. They showed three initial goal-directed interactions followed by one initial and one repeated non-targeted exploration. The sequences are similarly short as those in clusters 1 and 2 containing five interactions. These students show guessing behaviour that is partly goal-directed.

The comparison of the clusters regarding overall CPS skills using WLEs shows that all clusters have a negative average estimated skill (see Figure 10). Students engaging in unsystematic guessing or not answering (cluster 2) show the lowest average CPS skills, while students applying an incomplete approach (cluster 5) show the highest

TABLE 6 Comparison of mean CPS ability across clusters for correct responses in the climate control task

Compared clusters	Difference	Confidence interval		*p* value	Cohen's *d*
		Lower bound	Upper bound		
2–1	0.10	0.06	0.13	<.001	0.14
3–1	0.02	−0.01	0.05	.395	0.03
3–2	−0.08	−0.12	−0.04	<.001	−0.11

Abbreviation: CPS, complex problem-solving.

TABLE 7 Frequencies of clusters depending on item position for correct responses in climate control task

Cluster	Item position in test		χ^2	*df*	*p*	φ
	Early	Middle				
1	3,141	3,341	6.17	1	.013	0.03
2	1,400	1,540	6.67	1	.010	0.05
3	2,585	2,236	25.27	1	<.001	0.07

CPS skills in this group. CPS skills increase for clusters with longer sequences and with higher frequencies of goal-directed behaviour. The results of the Tukey Honest Significance Difference test show that significant differences in mean CPS skills exist between most clusters (see Table 8). Exceptions are clusters 4 and 1; and clusters 3, 5 and 6 showing similar CPS skills.

The comparison of the frequency of clusters between the early and the middle position in the test revealed that all guessing clusters (clusters 1, 2, 3 and 7) appeared more frequently at the early position (see Table 9). Also cluster 4 appeared more frequently at the early position. Cluster 6 is the only cluster that was observed more frequently at the middle position. For cluster 5, no significant difference was found.

3.2.2 | Correct solutions

We chose a solution with six clusters of sequences for correct solutions in the tickets task. The clusters are depicted in Figure 11. Cluster 1 contains the sequences of 2,629 students (21.77% of all correct solutions in the tickets task). The vast majority of the sequences showed initial goal-directed behaviour during the first three interactions followed by a peak of resets at interaction 4. This peak indicates that students navigated to the first ticket option and reset the task to consider further options. Starting from interaction 5, a lot of repeated goal-directed behaviour followed again by initial goal-directed behaviour took place. The sequences are of medium length with the longest sequence containing 21 interactions and an average sequence length of 9.29 interactions. The students apply a quite efficient (rather minimalistic) approach, since most of them show the minimum behaviour that is needed to solve the task.

Cluster 2 contains the sequences of 1978 students (16.38% of all correct solutions in the tickets task). The sequences show again almost exclusively initial goal-directed behaviour in the beginning followed by much repeated goal-directed behaviour. Only little non-targeted exploration took place. There are several peaks of resets at interaction 4, 9 and 13. The sequences are quite long with the longest sequence containing 27 interactions and an average sequence length of 20.99 interactions. These students seem to double-check the relevant tickets again and again.

Cluster 3 contains the sequences of 769 students (6.37% of all correct solutions in the tickets task). The sequences show again almost exclusively initial goal-directed behaviour in the beginning followed by repeated goal-directed behaviour. There is a peak of non-targeted exploration at interaction 5 and peaks of resets at interaction 4 and 6. Apart from the peak not much non-targeted exploration took place. The sequences are of medium length with the longest sequence containing 17 interactions and an average sequence length of 10.99 interactions. These students (as those in cluster 1) also show quite efficient behaviour, only they also show some non-targeted exploration.

Cluster 4 contains the sequences of 2,157 students (17.86% of all correct solutions in the tickets task). The sequences show again almost exclusively initial goal-directed behaviour in the beginning followed by much repeated goal-directed behaviour. Only little non-targeted exploration took place. There are peaks of resets at interaction 5 and 9. The sequences are rather long with the longest sequence containing 27 interactions and an average sequence length of 14.84 interactions. These students also show rather minimalistic behaviour (as those students in cluster 1). However, they investigated the tickets in a different order, forcing them to reset once more and navigate back to the ticket they inspected first.

Cluster 5 contains the sequences of 672 students (5.57% of all correct solutions in the tickets task). The sequences show exclusively initial goal-directed behaviour in the first three interactions followed by almost half non-targeted exploration and goal-directed behaviour. In the course of the task students exhibited more and more repeated interactions. In this group the reset button was not used at all. The sequences are quite short with the longest sequence containing 17 interactions and an average sequence length of 7.68 interactions. These students show guessing behaviour that is partly goal-directed and partly non-targeted exploration.

Cluster 6 contains the sequences of 3,869 students (32.04% of all correct solutions in the tickets task). The sequences show exclusively initial goal-directed behaviour in the beginning followed by repeated goal-directed behaviour. In this group neither the use of the reset button nor non-targeted exploration was displayed. All sequences are identical and contain 5.00 interactions. Therefore, these students show "goal-directed guessing".

The comparison of the clusters regarding CPS skills using WLEs shows that students in all clusters except one have a positive average

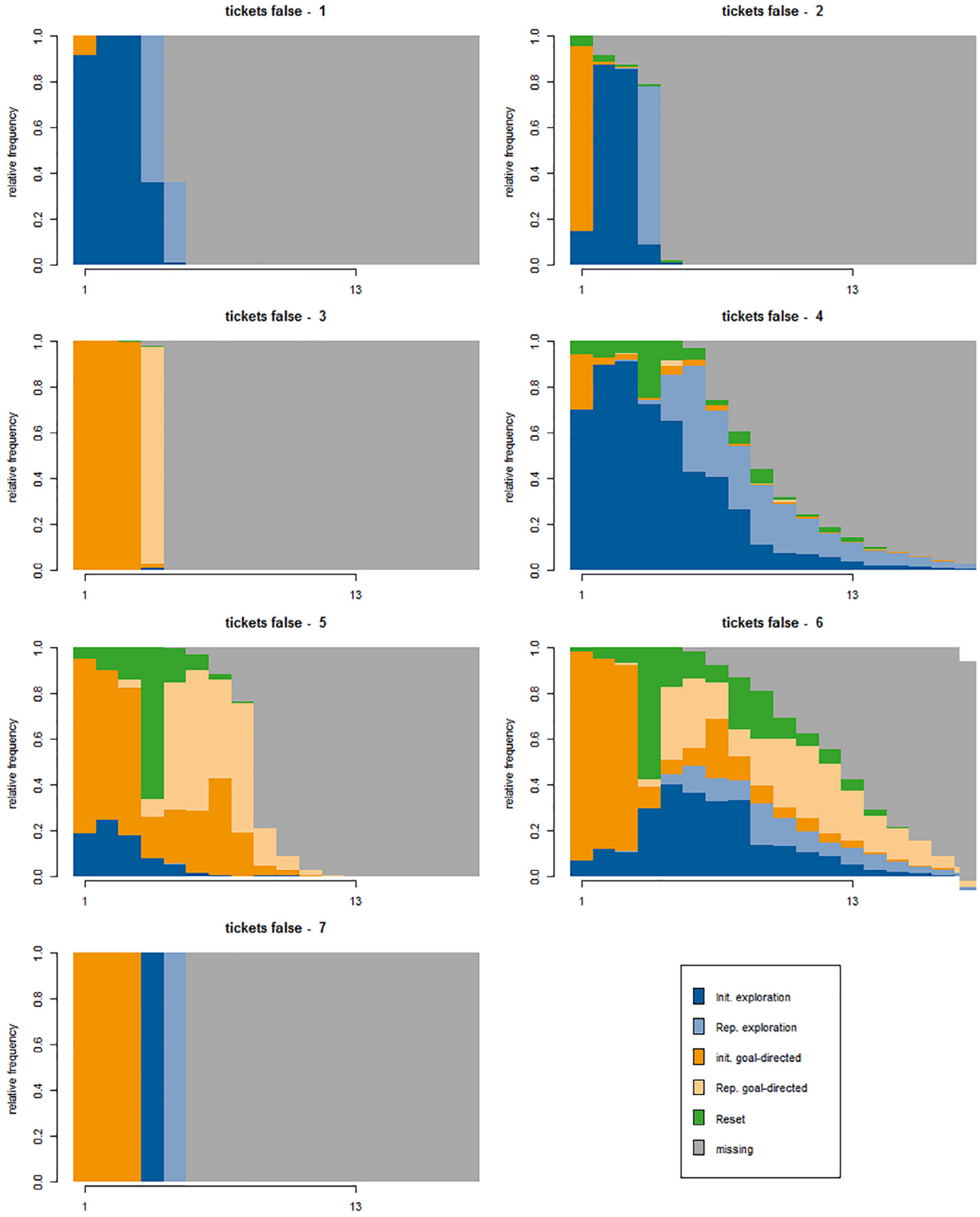

FIGURE 9 State distribution plots for each cluster for false solutions in the tickets task [Colour figure can be viewed at wileyonlinelibrary.com]

estimated skill (see Figure 12). Students, who double-checked their solution (cluster 2), show the highest average CPS skills, while students unsystematically guessing (cluster 5) show the lowest CPS skills. CPS skills increase for clusters with longer sequences and with higher frequencies of goal-directed behaviour. The results of the Tukey Honest Significance Difference test show that significant differences in mean CPS skills exist between all clusters (see Table 10).

The comparison of clusters with regard to their occurrence at different positions in the test revealed that the goal-directed guessing cluster (cluster 6) was observed more frequently at the early position (see Table 11). The minimalistic and the double-checking clusters (clusters 1, 2, 3 and 4) were observed more frequently at the middle position. For cluster 5, no significant difference was found.

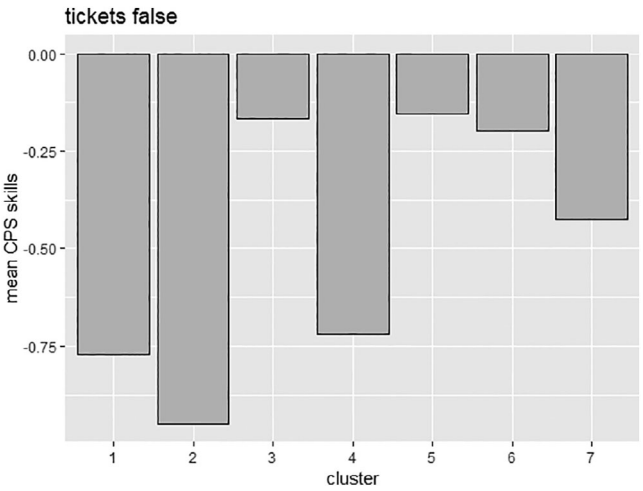

FIGURE 10 Mean CPS skills per cluster for false responses in the tickets task. CPS, complex problem-solving

4 | DISCUSSION

The aim of this study was to investigate how behavioural sequences are related to success or failure in CPS. Moreover, we wanted to clarify inconsistent findings of previous research regarding the usefulness of exploration in CPS. To this end, we used log data from the PISA 2012 CPS assessment and conducted full-path sequence analysis. We were able to clarify the inconsistent previous findings regarding exploration. Moreover, we identified several behavioural patterns associated with success or failure in CPS. Most patterns were found in both investigated tasks and thus across the two CPS frameworks of FSA and LSE. In the following paragraphs we will first discuss our results concerning non-targeted exploration and goal-directed behaviour (Hypotheses 1 and 2) before we discuss the behavioural patterns we found and their relation to students' overall CPS performance (Research questions 1 and 2).

4.1 | Non-targeted exploration and goal-directed behaviour

The results of the chi-squared test revealed that both initial and repeated non-targeted exploration were found more frequently among false responses. This finding supports our Hypothesis 1. Therefore, the results of Dormann and Frese (1994) and Eichmann et al. (2018), who reported a positive relation between exploration

TABLE 8 Comparison of mean CPS ability across clusters for false responses in the tickets task

Compared clusters	Difference	Confidence interval		*p* value	Cohen's *d*
		Lower bound	Upper bound		
2–1	−0.18	−0.24	−0.11	<.001	−0.17
3–1	0.60	0.55	0.65	<.001	0.89
4–1	0.05	−0.02	0.12	.428	0.11
5–1	0.61	0.54	0.69	<.001	0.98
6–1	0.57	0.51	0.63	<.001	0.95
7–1	0.35	0.26	0.44	<.001	0.56
3–2	0.78	0.72	0.84	<.001	1.00
4–2	0.23	0.15	0.31	<.001	0.27
5–2	0.79	0.71	0.87	<.001	1.10
6–2	0.75	0.68	0.82	<.001	1.10
7–2	0.52	0.43	0.62	<.001	0.70
4–3	−0.55	−0.62	−0.49	<.001	−0.77
5–3	0.01	−0.06	0.08	.999	−0.00
6–3	−0.03	−0.08	0.02	.533	0.01
7–3	−0.26	−0.34	−0.17	<.001	−0.38
5–4	0.56	0.48	0.65	<.001	0.84
6–4	0.52	0.45	0.60	<.001	0.81
7–4	0.30	0.20	0.40	<.001	0.43
6–5	−0.04	−0.12	0.03	0.648	0.01
7–5	−0.27	−0.37	−0.17	<0.001	−0.41
7–6	−0.23	−0.32	−0.13	<0.001	−0.40

Abbreviation: CPS, complex problem-solving.

TABLE 9 Frequencies of clusters depending on item position for false responses in tickets task

	Item position in test					
Cluster	Early	Middle	χ^2	*df*	*p*	φ
1	1,463	814	184.98	1	<.001	0.29
2	753	383	120.51	1	<.001	0.33
3	3,485	3,013	34.29	1	<.001	0.07
4	590	415	30.47	1	<.001	0.17
5	462	437	0.70	1	.404	0.03
6	851	973	8.16	1	.004	0.07
7	349	293	4.88	1	.027	0.09

and success in CPS, seem to be the result of not differentiating between goal-directed and non-targeted exploration. While goal-directed exploration might indeed be related to success, non-targeted exploration is in our data clearly related to failure. Therefore, non-targeted exploration could rather be a sign of confusion or distraction. These results were consistent in both investigated tasks.

We found that goal-directed behaviour (opposed to non-targeted exploration) was only if it was repeated and only in the tickets task more frequently found among correct responses. In the climate control task, this difference was not significant. We did not find significant differences for initial goal-directed behaviour in either task. Therefore, Hypothesis 2 was only supported for repeated goal-directed behaviour in the tickets task. A possible reason for that could be the higher opacity of the tickets task compared to the climate control task. While in the climate control task gathered information stays visible until the reset button is used, in the tickets task this is not the case. Therefore, in the tickets task repeated goal-directed behaviour might have been used to recall information, which was not required in the climate control task.

4.2 | Climate control task

4.2.1 | False solutions

We identified five clusters of behaviour sequences that did not result in correct solutions. Students in the cluster with the highest overall CPS skills among those who did not solve the climate control task showed an incomplete approach. They seem to have correctly applied the VOTAT strategy, but not to *every* input variable leading to an incomplete solution (cluster 5). Students that failed to apply the VOTAT strategy and instead investigated irrelevant content a lot showed medium overall CPS skills (clusters 2 and 4). Students in these two clusters seem to have applied a similar approach and differed mainly by sequence length. They engaged mainly in repeated non-targeted exploration behaviour, which could be a sign of over-estimating the relevance of in fact irrelevant information. Students mostly exploring irrelevant information and stopping their attempts early on showed even lower overall CPS skills (cluster 1). Since they show mainly initial non-targeted exploration, they do not seem to find any satisfying solution. However, the lowest overall CPS skills were shown by the group guessing an answer or leaving the task unanswered (cluster 3). These students' sequences were shorter than minimally required by the task. Therefore, they were assumed to be guessing.

4.2.2 | Correct solutions

We found three clusters of behaviour sequences resulting in correct solutions. Those students that seem to double-check their solutions showed the highest overall CPS skills among students working on the climate control task (cluster 2). Notably, only about 21% of the correct responses were in the double-checking cluster. There was no significant performance difference in overall CPS between students who applied a quite efficient approach (cluster 1) and students who applied a mixed approach of goal-directed behaviour and non-targeted exploration (cluster 3).

4.3 | Tickets task

4.3.1 | False solutions

We identified seven clusters of behaviour resulting in false solutions in the tickets task. Among those students who showed the highest overall CPS skills in the false solutions tickets group was one group that applied the incomplete approach of showing mostly goal-directed behaviour but failed to investigate all relevant tickets (cluster 5). But also those who showed a mixed approach of goal-directed behaviour and non-targeted exploration (cluster 6) and those who showed goal-directed guessing (cluster 3) were found among the highest performing group (of students who got the tickets task wrong). In the tickets task, all students who did not use the reset button, were assumed to be guessing, since they did not inspect all relevant tickets. Notably, the group of goal-directed guessers was by far the largest among the false responses in the tickets task (43.41%). Lower overall CPS skills were found among students who were also guessing but whose guesses were only partly goal-directed (cluster 7). They also seemed to apply goal-directed guessing, but in the end got

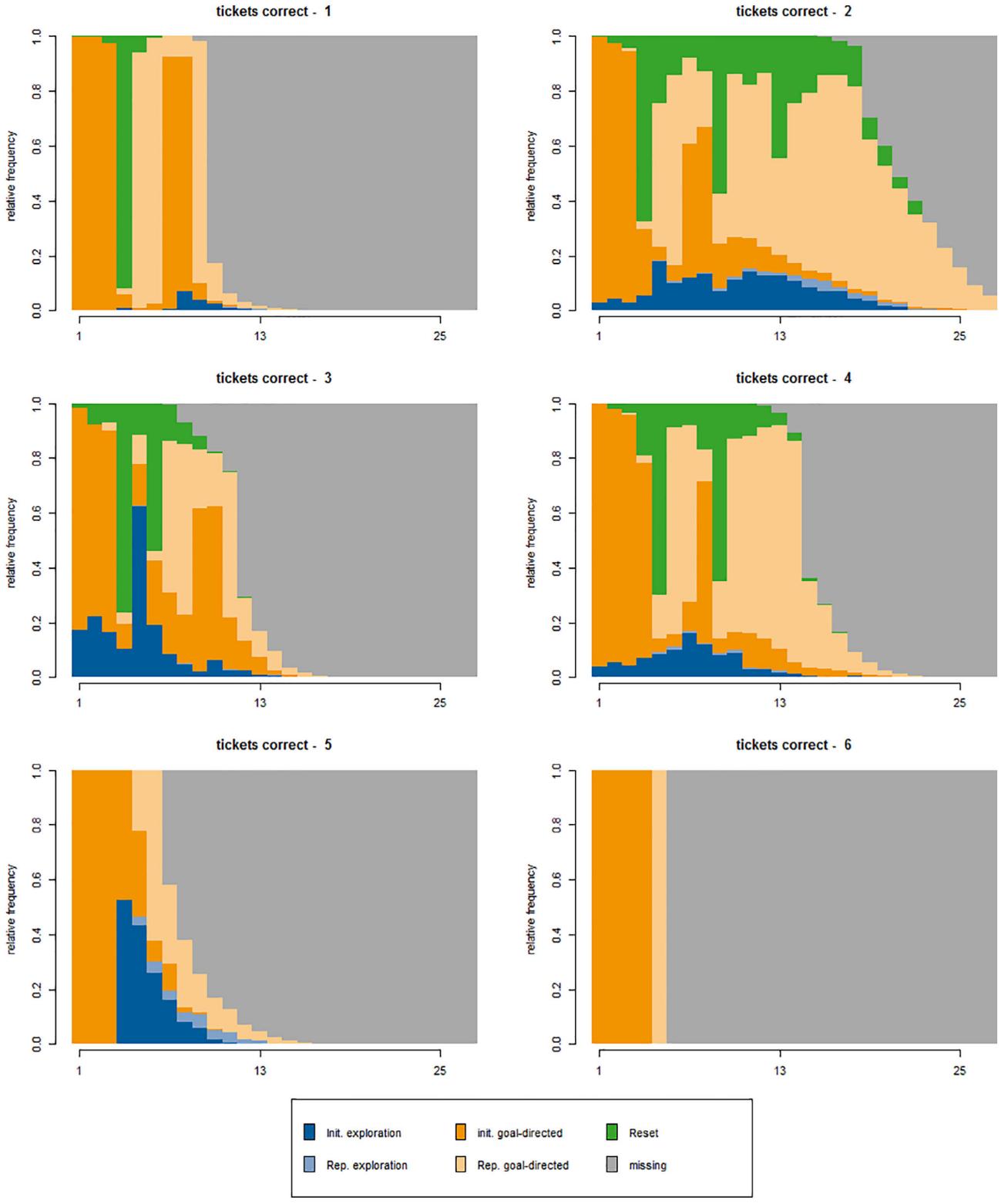

FIGURE 11 State distribution plots for each cluster for correct solutions in the tickets task [Colour figure can be viewed at wileyonlinelibrary.com]

sidetracked. An even lower overall CPS performance was found among those students who hardly showed any goal-directed behaviour. Within this group, it made no difference with regard to overall CPS performance whether students were guessing (cluster 1) or showed longer non-targeted exploration (cluster 4). However, the lowest overall CPS performance among those students who did not solve the tickets task was found with students who started goal-directed but then guessed an implausible solution or left the task

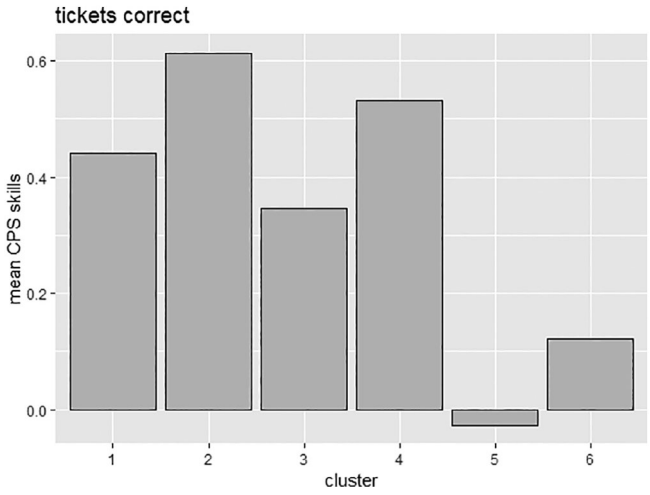

FIGURE 12 Mean CPS skills per cluster for correct responses in the tickets task. CPS, complex problem-solving

unanswered (cluster 2). In sum, 72.54% of the false solutions in the tickets task were the result of guessing (clusters 1, 2, 3 and 7). Therefore, guessing was the most frequent behaviour leading to false solutions in the tickets task.

4.3.2 | Correct solutions

We found six clusters of behaviour resulting in correct solutions in the tickets task. The highest overall CPS skills were again found among students who double-checked their solution (cluster 2). The second highest overall CPS skills were found among students who showed quite efficient behaviour in the tickets task (clusters 4 and 1). Those students, who were also efficient, but got distracted by non-relevant content at some point showed lower overall CPS skills (cluster 3). Even lower CPS skills were observed among students showing goal-directed guessing behaviour (cluster 6). The lowest skills of students correctly solving the

TABLE 10 Comparison of mean CPS ability across clusters for correct responses in the tickets task

		Confidence interval			
Compared clusters	Difference	Lower bound	Upper bound	*p* value	Cohen's *d*
2–1	0.17	0.12	0.23	<.001	0.26
3–1	−0.10	−0.17	−0.02	.005	−0.14
4–1	0.09	0.04	0.15	<.001	0.14
5–1	−0.47	−0.55	−0.39	<.001	−0.71
6–1	−0.32	−0.37	−0.27	<.001	−0.49
3–2	−0.27	−0.35	−0.19	<.001	−0.40
4–2	−0.08	−0.14	−0.02	.001	−0.12
5–2	−0.64	−0.72	−0.56	<.001	−0.96
6–2	−0.49	−0.54	−0.44	<.001	−0.75
4–3	0.19	0.11	0.27	<.001	0.27
5–3	−0.37	−0.47	−0.27	<.001	−0.53
6–3	−0.22	−0.30	−0.15	<.001	−0.34
5–4	−0.56	−0.64	−0.48	<.001	−0.82
6–4	−0.41	−0.46	−0.36	<.001	−0.62
6–5	0.15	0.07	0.23	<.001	0.23

Abbreviation: CPS, complex problem-solving.

TABLE 11 Frequencies of clusters depending on item position for correct responses in tickets task

	Item position in test					
Cluster	Early	Middle	χ^2	*df*	*p*	φ
1	1,152	1,432	30.34	1	<.001	0.11
2	777	1,158	75.02	1	<.001	0.20
3	336	417	8.71	1	.003	0.11
4	985	1,136	10.75	1	.001	0.07
5	338	319	0.55	1	.459	0.03
6	2,016	1,779	14.80	1	<.001	0.06

tickets task were shown by students who showed partly goal-directed and partly non-goal-directed guessing (cluster 5). In sum, the most frequent behaviour among the correct responses in the tickets task were rather minimalistic approaches (clusters 1, 3 and 4) with about 46%; the second most frequent were guessing approaches (clusters 5 and 6) with about 37.61%; and the least frequent was the double-checking approach (cluster 2) with 16.38%.

4.4 | General discussion

4.4.1 | Exploration behaviour

There are some characteristic behaviours that appear to be related to success in both tasks analysed in this study. In both tasks, the chi-squared tests revealed that non-targeted exploration is more frequent in false responses. This becomes also evident in the observed clusters, supporting the view of He and von Davier (2015) and Stadler et al. (2019), who found minimalistic behaviour to be most successful in CPS. However, only *repeated* goal-directed behaviour was found more frequently among correct responses and only in the tickets task. Moreover, since in both tasks the longest sequences were found among the correct responses and the shortest sequences were found among the false responses, sequence length seems to be positively related to success, in line with findings reported by Eichmann et al. (2018) and Naumann et al. (2014). Therefore, we assume long sequences of goal-directed behaviour, or in other words revisiting solution-relevant information, to be positively related to success while engagement with non-goal-directed information appears counterproductive. The distinction between goal-directed behaviour and non-targeted exploration we applied in the present study, therefore, reveals which specific types of behaviour are actually beneficial in CPS. What we defined as non-targeted exploration in our study seems to reflect confusion or distraction. In general, exploration is a rather broad category of behaviour that can include different things. The distinction of behaviours we applied, clarified the different roles exploration behaviour can fulfil in CPS. This finding has implications for the interpretation of past research outcomes as well as for the design of future research.

Our results help to clarify the ambiguous previous findings concerning the usefulness of exploration in CPS. Moreover, in line with our Research question we were able to identify patterns of behaviour that were associated with correct or false solutions in the CPS tasks analysed in this study and with overall CPS performance. We were also able to find commonalities and differences regarding these patterns between the tasks from two different CPS frameworks as intended with our Research question 2. We will further discuss the observed patterns in the following paragraphs.

4.4.2 | Perseverant approaches

In both tasks, we found students, who showed long sequences of goal-directed behaviour (i.e., who were presumably double-checking their solutions), to show the highest overall CPS skills. However, this was the least frequently used approach in both tasks. This double-checking approach could indicate these students' tendency to act perseverant or conscientious. This behaviour seems to be positively related to CPS performance, which is in line with the results of Naumann (2015), who found a positive effect of goal-directed actions in digital reading tasks, which likewise can be classified as complex. Therefore, perseverant goal-directed behaviour might be an adaptive strategy of dealing with complexity in general. However, at least in the climate control task we also found quite long sequences of mostly non-targeted exploration behaviour among the false responses. This, on the other hand, indicates that perseverance or conscientiousness does not necessarily lead to high CPS performance. Nevertheless, these "perseverant non-targeted explorers" still showed higher overall CPS performance than students, who exhibited shorter sequences of mostly non-targeted exploration behaviour, which again confirms the results of Naumann et al. (2014). Among those perseverant non-targeted explorers, there were groups of students showing more repeated non-targeted exploration while others showed more initial non-targeted exploration. Showing mostly initial non-targeted exploration might indicate that students identify much information but struggle to identify the relevant one. However, showing mostly repeated non-targeted exploration might indicate that students are convinced of an incorrect way to solve the problem. These behaviours, characterized by much non-targeted exploration, were more frequent in the climate control task. One possible reason is that acquiring knowledge through variable manipulation (as it is required in LSE tasks) might be quite uncommon to students and lead to behaviour of trial and error. Beckmann and Guthke (1995) argue that this kind of behaviour can be associated with high motivation and poor performance.

4.4.3 | Minimalistic approaches

However goal-directed perseverant approaches were quite successful, shorter sequences were more frequent in our data. Students' apparent preference of shorter sequences of behaviour might be due to the set time limit or students' limited motivation to engage in a low-stakes assessment such as PISA. However, it should be kept in mind that short sequences refer only to a small number of interactions and not to timing, which was not looked at in the present study. Efficient goal-directed behaviour (or minimalism) was found mostly among the correct responses. This finding is in line with the results of Stadler et al. (2019) who found minimalistic behaviour to be related to high CPS performance. However, the overall CPS skills of these students were lower than those applying double-checking behaviour. A reason for the lower overall CPS performance of the minimalists compared to the double-checking students could be that minimalists have a higher chance to oversee mistakes they made and therefore have a higher probability of giving a false response than students, who double-check their responses. Of course this interpretation implies that students exhibit similar behaviour across all CPS tasks.

4.4.4 | Guessing approaches

Even shorter sequences than the minimalistic ones were shown by students guessing a solution. In the tickets task, there was a remarkably large group of students guessing the correct solution. However, students in this group showed only medium overall CPS performance. Nearly one third of the correct solutions in the tickets task were the result of a goal-directed guessing approach. One reason why students applied this approach (quite successfully) mostly to the tickets and not to the climate control task might be that in the tickets task the solution required only one guess (i.e., buying one ticket) while in the climate control task guessing a solution would require independently guessing several relations between the variables decreasing the chance of guessing the correct solution. Therefore, in LSA tasks guessing might not be regarded as an adaptive strategy by students while it might be regarded as adaptive in FSA tasks, if one is not capable of solving the task properly. Especially, in scenarios like the FSA task we investigated, which was about buying a subway ticket, students might apply goal-directed guessing that does not guarantee an optimal but a sufficiently good solution: Instead of investing time and effort to find the cheapest ticket, students might choose to use a heuristic by buying any ticket that would satisfy their requirements (Evans, 2008; Gigerenzer & Goldstein, 1996). This behaviour could be an expression of either not being motivated to invest much time and effort or not being able to invest time and effort due to perceived time pressure or due to not having understood that there might be a better option available. In the context of a low-stakes assessment such as PISA a rather low motivation of students seems not surprising. In contrast, when buying real subway tickets, the motivation to save money might be higher. Although goal-directed guessing led to a high number of correct solutions in the tickets task, most goal-directed guessing led to incorrect (yet plausible) solutions. In a real situation these plausible solutions translate to not buying the cheapest but a valid ticket.

Opposed to goal-directed guessing, guessing randomly led mostly to false solutions in both tasks. Moreover, guessing or not answering seems to be the most frequent behaviour leading to false solutions in the tickets task. The difference between the goal-directed and the random guessers might be that goal-directed guessers read and understood the task (otherwise they could not identify goal-directed actions), whereas random guessers do not seem to have read and understood the task at hand. Random guessers also show a lower overall CPS performance than goal-directed guessers. Therefore, students applying random guessing seem to have more fundamental difficulties in CPS than students applying goal-directed guessing. The source for these difficulties could be (apart from lacking motivation), for example, low reading abilities, which prevent the students from properly understanding and processing the task, or struggling with different components of problem-solving (Carlson, Khoo, Yaure, & Schneider, 1990). Similarly to our result, Naumann et al. (2014) reported low achieving students in technology-based problem-solving to exhibit particularly little interactions with the tasks. Therefore, not engaging enough with problem-solving tasks might be one of the most frequent maladaptive behaviours.

4.4.5 | Incomplete approaches

Among the false solutions in both tasks there are also groups showing an incomplete goal-directed approach. These students started their process quite promising but failed to find the correct solution in the end. However, these students showed a relatively high overall CPS performance compared to other groups who did not solve the respective task. The overall CPS performance of this group in the tickets task was comparable to that of the goal-directed guessers. Since the incomplete goal-directed group's sequences are not especially short, these students do not seem to be particularly unmotivated. They seem to focus on plausible but wrong solutions. Repeating their goal-directed actions a lot, they do not seem perfectly convinced of their solution. However, they struggle with considering other options.

4.4.6 | Resetting

In both tasks, there seems to be more frequent use of the reset button in correct responses than in the false responses. This implies students' use of the reset button in both frameworks is indicative of a systematic approach rather than mere trial and error. Especially in the tickets task, which has a fixed sequence of actions following each other, an unsystematic approach becomes evident in the rare and unsystematic use of the reset button in false solutions, since this task requires the use of the reset button at certain points. Resetting might also reduce cognitive load. Especially in the climate control task, resetting clears the visible information of past actions, and therefore also reduces unnecessary information as potential sources of distraction (Sweller, 1988). According to Stadler et al. (2019), a reduction of cognitive load should be related to higher CPS performance. However, since generally little resetting was done, no final conclusions should be drawn with respect to resetting. Therefore, these results should be verified by future research.

4.4.7 | Effects of task position

We found most of the clusters either more frequently at the early or in the middle item position. The results were quite different for the two tasks. In the climate control task, shorter sequences were more often observed at the middle position, that is the item was presented at the beginning of the second half of the test. Longer sequences were observed more frequently at the early position. This finding could indicate a loss of students' motivation during the first half of the test. Greiff et al. (2018) reported a similar result in a latent class analysis investigating students' exploration behaviour in the course of six CPS tasks from the LSA framework. They argued that students who exhibited declining exploration probably experienced a decrease in motivation.

In the tickets task, however, the results were quite different. In this task, most of the guessing clusters were more frequently observed in the early item position, while the more successful,

minimalistic and perseverant approaches were observed more frequently in the middle position. This suggests a different process than the results regarding the climate control task. Since the tickets task was more difficult than the climate control task (fewer students were able to solve it correctly), the tickets task might have been too difficult at the beginning of the test. However, during the test students might have learned how to approach tasks from the FSM framework, so the tickets task was easier when presented at the middle position. Greiff et al. (2018) found that some students improve their exploration strategies during assessment leading to more elaborated task processing at later task positions. Since this pattern was only observed in the tickets task, students might learn adaptive strategies for processing FSM tasks more easily than strategies for LSA tasks. Another reason why tasks from both frameworks might differ with respect to strategy learning and motivation might be that tasks from the FSM framework can look quite heterogeneous whereas LSA tasks usually appear very similar. In the PISA 2012 test, all LSA tasks share surface features that make the tasks look quite similar, even if they concern different topics. Therefore, students who struggled with an LSA task before might be unmotivated when a second LSA task is administered to them. In contrast, in the PISA 2012 assessment FSA tasks varied for example with respect to response type and interface design. Therefore, students got the impression of rather heterogeneous tasks that possibly did not demotivate those students, who experienced difficulties before. On the contrary, it seems students acquired more adaptive strategies for processing FSM tasks during the assessment. Future research might address this issue by comparing students' behaviour in a larger number of tasks.

4.5 | Limitations

The present study used an exploratory and correlational approach to investigate behaviour in CPS. Therefore, the interpretation of our results needs to be validated by future research. Moreover, despite the fact that we used two tasks that may be seen as prototypical examples within their respective frameworks, the generalizability of our results will have to be established by analysing behaviour in a wider range of tasks. In addition, it should be investigated whether the results can also be replicated in more complex problem scenarios (e.g., systems such as those used by Stemmann & Lang, 2018). Further, our sample did only include 15-year-old students. Therefore, we cannot make assumptions about the behaviour in CPS in other age groups. Additionally, our large sample size limits the meaningfulness of the statistical significances we found to some extent. Another limitation is that we did not use a model-based approach such as latent class or profile analysis to identify groups of students. Therefore, the choice of our clustering solution relies on the comparison of relative quality criteria and not on model selection comparing goodness of fit measures (Oberski, 2016). Also the assumption that students exhibit similar behaviour across different CPS tasks needs to be further investigated. Moreover, it should be kept in mind that the analysis of log data does not allow to identify the intention that students had when exhibiting certain behaviours. Therefore, assumptions about causes of actions have to be validated in future research including experimental setups as well as think-aloud studies. Since PISA is a low-stakes assessment students' intentions might also be affected by a rather low motivation. Moreover, future research should also take timing information into account, since timing could be part of students' strategies in CPS. Further research is needed to overcome these limitations.

5 | CONCLUSION

We identified several behaviours associated with success or failure in CPS tasks. We observed a high proportion of goal-directed behaviour mostly among correct responses and a high proportion of non-targeted exploration mostly among false responses. Note that non-targeted exploration was defined as interactions *not necessary* to solve the task in this study, while required exploration was categorized as goal-directed behaviour. However, students applying double-checking approaches showed even higher CPS skills than students applying efficient, minimalistic approaches. Among the false solutions, extremely short behaviour sequences ultimately resulting in guessing a response are frequently observed especially in the tickets task. Therefore, the most frequent obstacles in CPS we found are abandoning a problem early and being sidetracked by goal-irrelevant content. Our findings hold true for both our investigated CPS tasks, however, the different behaviour patterns were found differently often in the two tasks. Thus, it seems our results are applicable to tasks from the LSE as well as from the FSA framework.

Overall, our results contribute to a better understanding of the processes that are related to success and failure in CPS. Moreover, they are a promising basis to make students more competent problem solvers. Encouraging students not to abandon problems early and teach them to identify and stick to the relevant aspects of problems might help them to become better problem solvers and prepare them for complex tasks they will most certainly encounter in their future. Moreover, knowledge about behaviour sequences related to success or failure in CPS makes it possible to identify the particular difficulties individual students are facing while solving complex problems. This information could be used to give students feedback about the aspects of their behaviour that are considered to be related to low CPS performance (Shute, 2008).

The detailed analysis of students' behaviour while solving complex problems allowed us to gain deeper insights into the processes in CPS. Most importantly, our results may help clarifying the role of exploration behaviour, specifically concerning the question of whether this kind of behaviour is beneficial in CPS. This knowledge may help to strengthen students' CPS skills and prepare them for the challenges of the 21st century.

ACKNOWLEDGEMENTS

This research was funded by the German Federal Ministry of Education and Research (Grant Numbers: 01LSA1504A and 01LSA1504B)

and by a project funded by the Fonds National de la Recherche Luxembourg (The Training of Complex Problem Solving; "TRIOPS").

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created in this study.

ORCID

Beate Eichmann https://orcid.org/0000-0001-7135-7945

REFERENCES

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, *14*, 471–494.

Apedoe, X. S., & Schunn, C. D. (2013). Strategies for success: Uncovering what makes students successful in design and learning. *Instructional Science*, *41*, 773–791. https://doi.org/10.1007/s11251-012-9251-4

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics*, *118*, 1279–1333. https://doi.org/10.1162/003355303322552801

Beckmann, J., & Guthke, J. (1995). Complex problem solving, intelligence, and learning ability. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 177–200). Hillsdale, NJ: L. Erlbaum Associates.

Bell, B. S., & Kozlowski, S. W. J. (2008). Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology*, *93*, 296–316. https://doi.org/10.1037/0021-9010.93.2.296

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht, the Netherlands: Springer. https://doi.org/10.1007/978-94-007-2324-5_2

Carlson, R. A., Khoo, B. H., Yaure, R. G., & Schneider, W. (1990). Acquisition of a problem-solving skill: Levels of organization and use of working memory. *Journal of Experimental Psychology: General*, *119*, 193–214.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.

Dormann, T., & Frese, M. (1994). Error training: Replication and the function of exploratory behavior. *International Journal of Human-Computer Interaction*, *6*, 365–372. https://doi.org/10.1080/10447319409526101

Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2018, April). *Using process data to explain group differences in complex problem solving*. Annual Conference of the the National Council on Measurement in Education (NCME), New York.

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education*, *128*, 1–12. https://doi.org/10.1016/j.compedu.2018.08.004

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Field, A., Miles, J., & Field, Z. (2013). *Discovering statistics using R (Reprint)*. Los Angeles, Calif.: Sage.

Frensch, P. A., & Funke, J. (1995). Definitions, traditions, and a general framework for understanding complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 3–22). Hillsdale, NJ: L. Erlbaum Associates.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning*, *7*, 69–89. https://doi.org/10.1080/13546780042000046

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, *40*. https://doi.org/10.18637/jss.v040.i04

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2014). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning*, *21*, 356–382. https://doi.org/10.1080/13546783.2014.989263

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, *126*, 248–263. https://doi.org/10.1016/j.compedu.2018.07.013

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36–46. https://doi.org/10.1016/j.chb.2016.02.095

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving. *Applied Psychological Measurement*, *36*, 189–213. https://doi.org/10.1177/0146621612439620

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with N-grams. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Springer proceedings in Mathematics & Statistics. Quantitative Psychology Research* (Vol. 140, pp. 173–190). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-19977-1_13

Jirout, J., & Zimmerman, C. (2015). Development of science process skills in the early childhood years. In K. C. Trundle & M. Saçkes (Eds.), *Research in early childhood science education* (pp. 143–165). Dordrecht, the Netherlands: Springer.

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log-data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*, 527–563.

Mislevy, R. J. (2019). On integrating psychometrics and learning analytics in complex assessments. In H. Jiao, R. W. Lissitz, & A. van Wie (Eds.), *Data analytics and psychometrics: Informing assessment practices* (pp. 1–52). Charlotte, NC: Information Age.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161.

Naumann, J. (2015). A model of online reading engagement: Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior*, *53*, 263–277. https://doi.org/10.1016/j.chb.2015.06.051

Naumann, J., Goldhammer, F., Rölke, H., & Stelter, A. (2014). Erfolgreiches Problemlösen in technologiebasierten Umgebungen: Wechselwirkungen zwischen Interaktionsschritten und Aufgabenanforderungen [Successful problem solving in technology rich environments: Interactions between number of actions and task demands]. *Zeitschrift für Pädagogische Psychologie*, *28*, 193–203. https://doi.org/10.1024/1010-0652/a000134

Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. In J. Robertson & M. Kaptein (Eds.), *Human-computer interaction series. Modern statistical methods for HCI* (1st ed., pp. 275–287). Cham, Switzerland: Springer.

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. *PISA*. Paris, France: OECD.

OECD. (2014a). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (Vol. V). Paris, France: OECD. https://doi.org/10.1787/9789264208070-en

OECD. (2014b). *PISA 2012: Technical report*. Paris, France: OECD. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical ComputingVienna, Austria. Retrieved from https://www.R-project.org/

Revelle, W. (2018). *Psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Richter, T., Naumann, J., & Noller, S. (2003). LOGPAT: A semi-automatic way to analyze hypertext navigation behavior. *Swiss Journal of Psychology*, *62*, 113–120. https://doi.org/10.1024//1421-0185.62.2.113

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test Analysis Modules*. Retrieved from https://CRAN.R-project.org/package=TAM

Schult, J., Stadler, M., Becker, N., Greiff, S., & Sparfeldt, J. R. (2017). Home alone: Complex problem solving performance benefits from individual online assessment. *Computers in Human Behavior*, *68*, 513–519. https://doi.org/10.1016/j.chb.2016.11.054

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153–189. https://doi.org/10.3102/0034654307313795

Signorell, Andri et mult. al. (2019). DescTools: Tools for descriptive statistics. Retrieved from https://cran.r-project.org/package=DescTools

Sonnleitner, P., Brunner, M., Keller, U., & Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *Journal of Educational Psychology*, *106*, 681–695. https://doi.org/10.1037/a0035506

Stadler, M., Fischer, F., & Greiff, S. (2019). Taking a closer look: An exploratory analysis of successful and unsuccessful strategy use in complex problems. *Frontiers in Psychology*, *10*, 248. https://doi.org/10.3389/fpsyg.2019.00777

Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior*, *65*, 100–106. https://doi.org/10.1016/j.chb.2016.08.025

Stemmann, J., & Lang, M. (2018). Eignet sich die logfilegenerierte Explorationsvollständigkeit als Prozessindikator für den Wissenserwerb im problemlösenden Umgang mit technischen Alltagsgeräten? [Is logfile-generated exploration completeness suitable as a process indicator for knowledge acquisition in handling of everyday technical devices?]. *Journal of Technical Education*, *6*, 185–199.

Studer, M. (2013). *WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R*. LIVES Working Papers, 24. doi: https://doi.org/10.12682/lives.2296-1658.2013.24.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *179*, 481–511. https://doi.org/10.1111/rssa.12125

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research*, *40*, 471–510. https://doi.org/10.1177/0049124111415372

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285. https://doi.org/10.1207/s15516709cog1202_4

Tóth, K., Rölke, H., Greiff, S., & Wüstenberg, S. (2014). Discovering Students' Complex Problem Solving Strategies in Educational Assessment. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Chairs), Proceedings of the 7th International Conference on Educational Data Mining London : CEUR Workshop Proceedings. Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/short%20papers/225_EDM-2014-Short.pdf

Wilpert, B. (2009). Impact of globalization on human work. *Safety Science*, *47*, 727–732. https://doi.org/10.1016/j.ssci.2008.01.014

Wirth, J. (2004). *Selbstregulation von Lernprozessen* [Self-regulation of learning processes]. *Pädagogische Psychologie und Entwicklungspsychologie* (Vol. 39). Münster, Germany: Waxmann.

Wüstenberg, S., Greiff, S., Molnár, G., & Funke, J. (2014). Cross-national gender differences in complex problem solving and their determinants. *Learning and Individual Differences*, *29*, 18–29. https://doi.org/10.1016/j.lindif.2013.10.006

How to cite this article: Eichmann B, Greiff S, Naumann J, Brandhuber L, Goldhammer F. Exploring behavioural patterns during complex problem-solving. *J Comput Assist Learn*. 2020; 1–24. https://doi.org/10.1111/jcal.12451