Leibniz Universität Hannover

Fakultät für Mathematik und Physik

Institut für Radioökologie und Strahlenschutz

Masterarbeit

EXIT-Spiele im Physikunterricht: Eine empirische Analyse über den Kompetenzzuwachs der Schüler*innen

---

vorgelegt von: Onno Maximilian Rüther

Matrikelnummer: 10014491

Abgabedatum: 17.08.2022

Erstprüfer: Prof. Dr. Clemens Walther

Zweitprüfer: Dr. Jan-Willem Vahlbruch

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

1. Einleitung

Viele Jahre vor dem Aufkommen der ersten ESCAPE-Räume und EXIT-Spiele respektive einem unterrichtlichen Einsatz derselben erkannte der deutsche Pädagoge Menzel den positiven Einfluss, den das Spielen in der Lehre generell auf die Lernprozesse von Schüler*innen hat, wie das nachfolgende Zitat aussagekräftig belegt. *„Wer in der Schule nicht spielen lernt, lernt nicht lernen.“* [1]

Mittlerweile sind die Effekte des Spielens auf die Motivation von Schüler*innen und die Lernprozesse selbst vielfältig empirisch untersucht und erklärt. Die pädagogische Psychologie hat festgestellt, dass Spielen der Elaboration und Organisation von Wissen dienlich ist, sodass bereits bestehende Strukturen und Fakten miteinander vernetzt und gefestigt werden können. Dies sorgt schließlich für einen Übergang dieser Strukturen aus dem Kurzzeitgedächtnis in das Langzeitgedächtnis [2]. Für eine gelingende Elaboration durch Spiele ist die Kompetenz des Problemlösens von essentieller Bedeutung [3]. Kipman beschreibt das spielerische Lernen durch Problemlösen als einen unterbewussten Prozess indem Schüler*innen sich immer wieder an neuen Herausforderungen messen, mit ihrer Strategie scheitern, um dann mit einer neuen Herangehensweise erneut eine Lösung des Problems herbeiführen wollen [4]. Die entscheidende Rolle des Lernens aus Fehlern auch im spielerischen Kontext muss laut Heinecke vor der speziellen Fehlerkultur im Physikunterricht betrachtet werden, die sie mit Aussagen wie: *„Physik ist das was nie gelingt.“* [5], *„Physik habe ich nie verstanden.“* [5], oder *„In Physik sage ich lieber nichts - das ist ja doch meistens falsch.“* [5] untermauert [5]. Dieser inadäquate Umgang mit Fehlern wird auch an der erst kürzlichen Wandlung des lange gebräuchlichen Begriffs „Fehlvorstellungen“ zu Schüler*innenvorstellungen deutlich [5]. An dieser Stelle können EXIT-Spiele ansetzten. Durch die gezielte Herausforderung der Schüler*innen in kooperativen Lernstrukturen kann eine Umdeutung der vorherrschenden Leistungssituation der Schule, in der Fehler sanktioniert werden, zu einer fehlertoleranten Lernsituation erreicht werden [5, 6]. Neben dieser Umdeutung bieten die meisten Spiele für den Unterrichtseinsatz die Möglichkeit, bestehendes Wissen zu wiederholen, zu transferieren, und zu vernetzen [7]. Der Einfluss dieses Übens sollte laut Hepp nicht unterschätzt werden und ist für ein physikalisches Verständnis der Schüler*innen maßgeblich. Physikalische Phänomene wirkten ansonsten auf diese nur wie ein inhaltsloses Wortgerüst [8].

Mit dem in einer vorangegangenen Arbeit entwickelten Methodenwerkzeug des EXIT-Spiels ist ein Produkt entstanden, das alle zuvor beschriebenen Vorteile in sich vereint [9]. Jedoch stehen die meisten Lehrer*innen dem Einsatz von neuen Unterrichtskonzeptionen kritisch gegenüber [10]. Daher binden viele Lehrer*innen diese nur sehr zurückhaltend in ihren Unterricht ein, weshalb eine empirische Untersuchung der Lernwirksamkeit sinnvoll erscheint [10]. Weiterhin ist die Wirksamkeit von EXIT-Spielen für den Schulunterricht bislang nicht ausreichend untersucht [11].

Anhand dieses Streifzugs durch die Welt des spielerischen Lernens lässt sich der Bedarf an empirischen Untersuchungen zum Spieleinsatz im Unterricht erahnen. Die Vielzahl an angeschnittenen Themenbereichen verdeutlicht die Komplexität, die diese Methode zur Gestaltung von Unterricht mit sich bringt. Damit einher geht die Frage, ob und wenn ja in wieweit konkret das Spielen im Physikunterricht eine lernförderliche Wirkung bei den Schüler*innen hervorrufen kann. Aus diesem Grund wird in der nachfolgenden Arbeit, im Sinne der Lernprozessdiagnostik aus der pädagogischen Psychologie [12], mittels eines quasiexperimentellen Pre-Posttest-Designs der Kompetenzzuwachs durch den Einsatz des erwähnten EXIT-Spiels untersucht. Anhand von vier Klassen der Sekundarstufe I soll weiterhin über den Vergleich von Experimental- und Kontrollgruppen die Lernwirksamkeit des EXIT-Spiels gegenüber konventionellen Unterrichtskonzeptionen zur Radioaktivität und Kernphysik untersucht werden.

2. Theoretischer Hintergrund

2.1. Spielerisches Lernen

Der Grund und die Funktion des Spielens beschäftigten die Menschen schon seit vielen hundert Jahren. Daher existiert auch eine Vielzahl an Spieltheorien, die unterschiedliche Erklärungsversuche des Spielens darstellen. Schon vor Beginn des 20. Jahrhunderts setzten sich große Psychologen wie Spencer, Lazarus und Groos mit Spielen auseinander [13]. Letzterer sieht im Spielen eine Art Vorübungsfunktion zur Bewältigung des Lebens im Erwachsenenalter [13]. Im 20. Jahrhundert prägten bekannte Persönlichkeiten wie Freud, Erikson und Piaget die Spieltheorie [11]. Piaget bedient sich der grundlegenden Idee von Groos, entwickelt diese allerdings weiter. So versteht er im Spielen eine Übung der aktuellen Intelligenz [13]. Kindliches Spielen stellt für Piaget einen Weg zur Erkenntnis der Wirklichkeit dar, da Spiele immer in Auseinandersetzung mit der Umwelt stattfinden [2]. Ferner dienten Spiele der Festigung von bereits erworbenem Wissen [13]. Es besteht also die Möglichkeit, dass Schüler*innen ihr Wissen durch den Einsatz von EXIT-Spielen festigen können. Dabei ist der intendierte Zweck des Spiels oftmals nicht deckungsgleich mit der Wahrnehmung der Schüler*innen, der Lernprozess erfolgt unterbewusst [3]. Huizinga definiert in seinem Buch „Homo Ludens – der spielende Mensch" das Spiel wie folgt: *„Spiel ist eine freiwillige Handlung oder Beschäftigung, die innerhalb gewisser festgesetzter Grenzen von Zeit und Raum nach freiwillig angenommenen, aber unbedingt bindenden Regeln verrichtet wird, ihr Ziel in sich selbst hat und begleitet wird von dem Gefühl der Spannung und Freude und einem Bewusstsein des Andersseins als das gewöhnliche Leben."* [14]

Tabelle 1: Merkmale des Spielens nach Meyer [15].

	Merkmal
1	Spielen erfordert einen freien Raum, da es selbst frei von fremden Zwecken ist.
2	Spielen ist in sich zielgerichtet.
3	Spielen findet in einer Scheinwelt statt.
4	Spielabläufe sind mehrdeutig und offen.
5	Spielen schafft eine handelnde Auseinandersetzung mit Spielenden oder dem Spielobjekt.
6	Spielen erfordert die Anerkennung von Spielregeln.
7	Im Spielen müssen gleiche Rechte und Gewinn- oder Beteiligungschancen für alle Mitspieler*innen bestehen.
8	Spiele erfüllen sich in der Gegenwart.
9	Spielen macht Spaß.

Aus dieser Definition gehen einige gemeinsame Merkmale von Spielen hervor, die Meyer, wie in Tabelle 1 dargestellt, zusammengefasst hat. Diese neun Merkmale dienten in der Entwicklungsphase des genutzten Spiels als Ankerpunkte, die in die Planung der einzelnen Rätsel und des gesamten Spiels mit einbezogen wurden [9]. Jedoch sollten besonders im Kontext Schule nicht alle Merkmale als passgenau angesehen werden. Kritisch zu betrachten ist das erste Merkmal, da das Spielen in der Lehre nicht zweckfrei sein kann, wie das nachfolgende Zitat von Meyer belegt: *„Spielen im Unterricht ist nicht zweckfrei, sondern ein zielgerichteter Versuch zur Entwicklung der sozialen, kreativen, intellektuellen und ästhetischen Komponenten der Schüler."* [15] Mikelskis-Seifert und Berendt untermauern diese These Meyers durch folgende Aussage: *„Die Grundidee des Spielens mit dem Ziel des Lernens ist es, die Motivation des Spielens zur Aneignung von Wissen einzusetzen."* [14] Daher kann davon ausgegangen werden, dass Spiele im (Physik-) Unterricht eine didaktische Legitimation haben, da höhere Ziele verfolgt werden können. Auf dieses Potential des Spielens für den Physikunterricht gehen Mikelkis-Seifert und Berendt ebenfalls ein. Hierbei beziehen sie sich auf Meyer, der ebenso wie Fürstenau zusammenfasst, dass Spielen das ganzheitliche Lernen, die Selbsttätigkeit der Schüler*innen, soziale Erfahrungen, die Klassengemeinschaft, sowie das Anwenden und Vertiefen des gelernten Wissens fördern kann [16]. Auer stellt als wesentlichen Vorteil des Spielens die Möglichkeit der Wahrnehmung von Spaß im Physikunterricht fest, wie das nachfolgende Zitat anschaulich belegt. *„Der Einsatz von verschiedenen Spielen im Physikunterricht ist eine Gelegenheit, für die Schüler eine anregende und interessante Lernumgebung zu schaffen, bei der sie die Physik mit Spaß verbinden können."* [17] Auf diese Vorteile des Spielens in der Lehre wird in nachfolgenden Kapiteln der Arbeit immer wieder vertiefend eingegangen.

Weiterhin gelingt es, durch den Einsatz des Mediums Spiel vom lehrerzentrierten Unterricht abzukehren und den Lehrer*innen mehr Zeit zur Verfügung zu stellen, damit diese von der diagnostischen Funktion des Spielens, bezogen auf den Lernerfolg, profitieren können [18]. Auch der „Huckepackeffekt", der besagt Lernen erfolge oftmals beiläufig, untermauert Mikelkis-Seiferts These, dass Spielen für den Lernprozess förderlich ist [14]. Aus den hier geschilderten Vorteilen wird ersichtlich, dass vom Einsatz des Mediums Spiel nicht nur die Schüler*innen, sondern auch die Lehrer*innen profitieren können, wobei sich für Lehrkräfte ein Mehraufwand an Vorbereitungsarbeit einstellt [17]. Zum besagten Mehraufwand gehört vor allem die ausführliche Planung der Unterrichtsstunde,

in der gespielt werden soll. Meyer bezeichnet diese als „Spielstunden“ [15]. Um Lehrer*innen in der Vorbereitung von Spielstunden zu unterstützen, hat Meyer eine aus sieben Fragen bestehende Checkliste entwickelt, mit welcher Lehrer*innen eine erfolgreiche Spielstunde planen können [15].

Tabelle 2: Checkliste zur erfolgreichen Durchführung einer Spielstunde nach Meyer [15].

	Element
1.	Warum will ich mit meinen Schüler*innen spielen?
2.	Welche Interessen könnten die Schüler*innen am Spiel haben?
3.	Welche Vorkenntnisse und Erfahrungen können die Schüler*innen einbringen?
4.	Wie lauten die Spielregeln?
5.	Wer ist Spielleiter?
6.	Müssen die Spielgruppen vor Spielbeginn bestimmt werden?
7.	Welche Spielmaterialien, Geräte oder Requisiten müssen besorgt werden?

Für die Vorbereitung einer Spielstunde, in der ein EXIT-Spiel gespielt werden soll, sind die Fragen fünf und sechs von nachrangiger Bedeutung. Dafür sollten sich Lehrer*innen besonders mit den Fragen zwei, drei und sieben auseinandersetzen [9, 19]. Zusätzlich zu Meyers Hilfestellung bei der Gestaltung von Spielstunden wird im Kapitel 5.1 beschrieben, wie der Vorbereitungsaufwand für den Einsatz des Spiels durch besondere Schwerpunktsetzung in der Entwicklung weiter reduziert werden konnte und auch notwendige Anpassungen möglichst effizient gestaltet wurden.

Abschließend lässt sich also festhalten, dass der Einsatz von Spielen im Physikunterricht Vorteile hat, wie die Feststellungen von Meyer, Mikelskis-Seifert, Behrendt und Auer zeigen. Weiterhin dient das Spiel nach Piaget, der Festigung von bereits erlerntem Wissen [13] und ist neben der reinen Vermittlung von Wissen auch der Motivation von Schüler*innen dienlich [14, 17, 20].

2.2. Entstehung und Gestaltungsmerkmale von ESCAPE-Räumen und EXIT-Spielen

ESCAPE-Räume und deren Brettspielderivate (EXIT-Spiele) haben seit Beginn der 2000er Jahre immer mehr an Bedeutung als Unterhaltungsaktivität gewonnen und haben in den letzten Jahren auch Einzug in die schulische und universitäre Bildung erhalten [11, 21, 22]. Wiemker definiert ESCAPE-Räume als eine Art Spiel in dem eine Gruppe in einer vorgegebenen Zeitspanne durch Lösen von Rätseln aus einem Raum entkommen

muss [22]. Zu einer ähnlichen Definition gelangt auch Veldkamp, die jedoch zusätzlich den Live-Action-Charakter der Spiele hervorhebt [21].

Der Ursprung des Spielgenres geht auf computerbasierte Spiele in den 80er und 90er Jahren des vergangenen Jahrtausends zurück [11]. Die ersten Live-Action ESCAPE-Räume entwickelten sich ab 2007 in Japan und gelangten einige Jahre später nach Europa [11, 23]. Im Verlauf der letzten Jahre haben ESCAPE-Räume immer mehr an Popularität gewonnen. So hat sich allein die Zahl der weltweiten ESCAPE-Räume von ca. 2800 im Jahr 2015 auf über 7200 im Jahr 2018 gesteigert [24].

Der generelle Aufbau eines solchen ESCAPE-Raums wird bereits grundlegend durch die Definition beschrieben. Die Spielenden sind in einem Raum eingesperrt und müssen durch das Lösen mehrerer Rätsel dem Raum in einer vorgegebenen Zeit entkommen. Wiemker beschreibt dies als die „*klassische dreigliedrige Spielschlaufe"* [22] des Escape-Raums. Es gibt eine Herausforderung aus der eine Lösung resultiert, die am Ende zu einer Belohnung für das Überwinden der Herausforderung führt [22]. Ein klassisches Beispiel hierfür, welches auch für diese Untersuchung verwendet wird, stellt ein verschlossener Gegenstand dar, der mithilfe eines Codes geöffnet werden kann. In diesem Gegenstand befindet sich eine Belohnung in Form weiteren Rätselmaterials. Zwei wesentliche Kategorien von Rätseln lassen sich dabei unterscheiden. Während mentale Rätsel kognitive Herausforderungen darstellen, die logisches Denken erfordern, versteht Wiemker unter physischen Rätseln Aufgaben, die die physische Bearbeitung von Gegenständen erfordern und vorwiegend dem Zeitverbrauch dienen [22]. Für das der Arbeit zugrunde liegende EXIT-Spiel sind insbesondere die mentalen Rätsel von zentraler Bedeutung, da sie für den antizipierten Lerngewinn verantwortlich sind [22].

Neben einer Kategorisierung einzelner Rätseltypen gilt es weiterhin, den Aufbau der Rätselstruktur durch das gesamte Spiel zu unterscheiden. Typischerweise bestehen solche Spiele nicht nur aus einem einzelnen, sondern mehreren, oft aufeinander aufbauenden Rätseln. Dabei lassen sich drei wesentliche Designformen unterscheiden, die in Abbildung 1 dargestellt werden. In der offenen Struktur können mehrere Rätsel zur selben Zeit bearbeitet werden, jedoch werden alle benötigt, um das finale Rätsel zu lösen (s. Abb. 1, A). Mit der sequentiellen Struktur, die im Spiel dieser Untersuchung vorlag, wird den Spielenden und Entwickelnden eine möglichst simple Struktur vorgebeben, in welcher jedes Rätsel zum nachfolgenden Rätsel führt und alle in der vorgegebenen Reihenfolge bis hin zum finalen Rätsel bearbeitete werden müssen (s. Abb. 1, B). Die pfadbasierte Struktur stellt eine Kombination beider Designformen dar (s. Abb. 1, C). Es gibt mehrere

Rätsel die gleichzeitig bearbeitet werden können und dennoch werden Informationen aus vorherigen Rätseln benötigt, um im Spiel voranzukommen. Es sind auch weitere, sehr komplexe Rätselstrukturen denkbar, die an mehreren Stellen eine Kombination der genannten Möglichkeiten darstellen, als Beispiel kann hier die Pyramidenstruktur genannt werden (s. Abb. 1, D) [21].

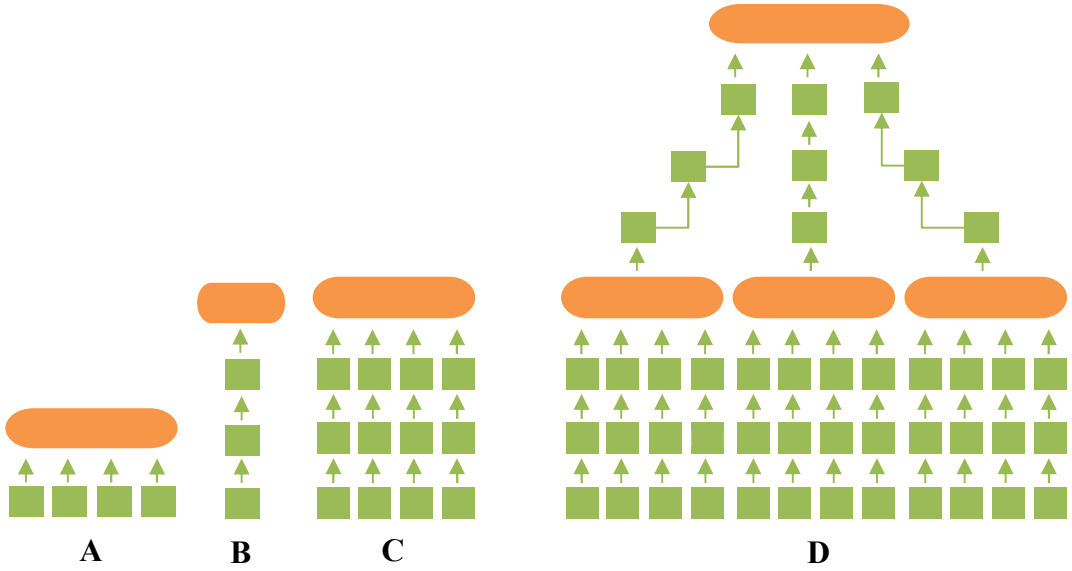


Abbildung 1: Rätselstruktur in ESCAPE-Räumen [21].
Bemerkung: **A** offen, **B** sequentiell, **C** pfadbasiert, **D** komplexe Pyramiden Struktur. Grüne Rechtecke sind Rätsel, orangene Ellipsen sind erreichte Zwischen- bzw. Endziele.

Die beschriebenen Rätsel sind in eine Rahmenhandlung oder Geschichte eingebunden, die den Spielenden eine Art der Immersion, also ein Eintauchen in eine virtuelle Realität, ermöglicht. Zusammen mit der Gestaltung des Raums oder Spielmaterials ist die virtuelle Realität maßgeblich für das „Flow-Erlebnis“ verantwortlich [11]. Unter dem erstmals von Csíkszentmihályi beschriebenen „Flow-Erlebnis“ wird ein Zustand verstanden, in dem sich Spielende vollständig auf die zu lösende Aufgabe konzentrieren [22]. In extremen Fällen reicht dieser Zustand soweit, dass Spielende alltägliche Routinen vergessen [22]. Besagtes „Flow-Erleben“ tritt jedoch nur ein, wenn das Spiel eine ausgewogene Balance zwischen Kompetenz der Spielenden und Schwierigkeitsgrad der Rätsel vorweisen kann. Nur dann sorgt das Flow-Erleben für eine kontinuierliche Aufrechterhaltung der Aufmerksamkeit und Motivation während des gesamten Spiels [11].
Die meisten der beschriebenen Eigenschaften und Gestaltungsmerkmale von ESCAPE-Räumen treffen auch auf die einige Zeit später erschienenen EXIT-Spiele zu und damit

auch auf das in dieser Untersuchung verwendete Spiel. Der größte und gleichzeitig wichtigste Unterschied hierbei ist die nicht vorhandene räumliche Gestaltung. Die komplette Immersion und damit auch das „Flow-Erlebnis“ hängen an der Geschichte und dem Spielmaterial, das zur Verfügung gestellt wird.

2.3. Einsatz von ESCAPE-Räumen und EXIT-Spielen als Methodenwerkzeug im Unterricht

Nachdem bereits im vorangegangenen Kapitel dargestellten Entwicklungsprozess, den ESCAPE-Räume in den letzten Jahren zu einem populären Unterhaltungsinstrument durchlaufen haben, wurde deren Potential für Bildungszwecke entdeckt. Dies führte zu einer verstärkten Integration insbesondere von Brettspielderivaten an schulischen und universitären Einrichtungen [18, 24]. O`Brien unterstützt dieses Vorgehen und ist der Meinung, dass unterrichtliches Spielen eine angemessene Möglichkeit darstellt, um das bisherige inhaltsbezogene und auf Instruktion basierende Bildungssystem hin zu einem schüler*innenzentrierten und auf das lebenslange Lernen ausgerichtete System zu wandeln [25]. Die Spiele wurden jedoch nicht, wie von deren Entwicklern ursprünglich intendiert, nur zur Unterhaltung in den Unterricht eingebunden, sondern verfolgen, wie Meyer beschreibt, immer von den Lehrpersonen erdachte Lernziele [15].

Tabelle 3: Unterschiede zwischen konventionellen Escape Räumen und Escape Räumen zum Einsatz im Unterricht. [18].

	Konventionelle ESCAPE-Räume	**ESCAPE-Räume im Unterricht**
Spielende	breit gefächert	spezielle Zielgruppe mit definierten Lernzielen
Erfolgsrate	variabel	hoch
Rätsel	keine Curriculums Abstimmung	auf das Curriculum abgestimmt
Rätselergebnisse	variabel	numerische oder alphabetische Codes
Spielorte	einer oder mehrere verbundene Räume	limitiert (Klassenraum)
Zeitbedarf	freie Zeitgestaltung	limitiert (Stundenplan)
Teilnehmerzahl	3-7	vollständige Lerngruppe (20-30) aufgeteilt in Kleingruppen (4-6)

An dieser Stelle setzt auch Veldkamp an und fasst die wesentlichen Unterschiede zwischen ESCAPE-Räumen zum Vergnügen und deren Einsatz im Unterricht zusammen (s. Tabelle 3). Der lernzielorientierte Einsatz von Spielen wird vor allem durch die von den

Spielenden abverlangten Kompetenzen sichergestellt. Exemplarisch seien hier das Problemlösen und Arbeiten im Team genannt, auch wenn noch viele weitere Kompetenzen während des Spielens benötigt und gefördert werden [11, 22]. Die förderbaren Kompetenzen resultieren direkt aus den bereits beschriebenen Rätselstrukturen und der Arbeit in einer kleinen Gruppe von Schüler*innen. Einschlägige Untersuchungsergebnisse belegen hier eine optimale Gruppengröße von vier bis sieben Spielenden pro Gruppe [18, 26, 27]. Auf die Rolle der Gruppengröße wird in einem weiteren Kapitel dieser Arbeit ausführlicher eingegangen.

Sowohl von Lehrer*innen als auch Schüler*innen werden ESCAPE-Räume und auch auf dem Konzept basierende Spiele als angemessenes Mittel zur Wiederholung, Übung und Vertiefung von bestehendem Wissen und Kompetenzen wahrgenommen [19]. Neben der reinen Betrachtung des Kompetenzerwerbs ist vor allem die gesteigerte Motivation der spielenden Schüler*innen, im Vergleich zu konventionellen Unterrichtskonzeptionen, ein primäres Ziel des Spieleinsatzes im Unterricht. Diese beeinflusst erfolgreiches Lernen maßgeblich [11, 28]. Auch wenn die Effektivität des Einsatzes von ESCAPE-Räumen und EXIT-Spielen im Unterricht noch nicht ausreichend untersucht worden ist, lassen erste Studienergebnisse einen positiven Effekt über breite Anwendungsbereiche und Altersstufen beobachten [11, 29, 30]. Aus diesen ersten Untersuchungen sind neben den positiven Auswirkungen auf die Kompetenz der Schüler*innen auch eine Reihe von weiteren Vorteilen, sowie Hindernissen des Spieleinsatzes belegt worden. Zu den Vorteilen zählt neben der gezielten Schulung einer Vielzahl der im Kerncurriculum vorgegebenen prozessbezogenen Kompetenzen, dass die Schüler*innen miteinander kommunizieren, analysieren, kritisch denken und durch die Immersion im Spiel die Motivation für die Unterrichtsinhalte erleichtert wird [11, 31]. Zu den größten Hindernissen des Spieleinsatzes gehören der große Zeitaufwand zur Implementierung von Spielen im Unterricht, der große Material- und Kostenaufwand für die Erstellung der Spiele und besonders die Gruppengröße [11]. Vor allem die mit der Entwicklung und Erprobung der Spiele verbundenen Schwierigkeiten sind für tertiäre Bildungseinrichtungen wie Universitäten von nachrangiger Bedeutung, da diese oftmals besseren Zugang zu finanziellen und zeitlichen Ressourcen haben als sekundäre Bildungseinrichtungen, wie weiterführende Schulen. Aus diesem Grund findet sich die Großzahl der entwickelten und auch regelmäßig eingesetzten Spiele an Universitäten [26].

Trotz der Hindernisse, die der Einsatz von ESCAPE-Räumen und EXIT-Spielen im Unterricht mit sich bringt, nehmen sowohl Schüler*innen als auch Lehrer*innen den Einsatz

dieser als positiv wahr [19]. Der hohe Zeitaufwand wird akzeptiert, da Vorteile des Einsatzes, wie die merklich gesteigerte Motivation, überwiegen [19]. Aufgrund des steigenden Bekanntheitsgrades, der positiven Wahrnehmung und Lerneffekte hat sich im Laufe der letzten Jahre ein größer werdendes Spektrum von ESCAPE-Räumen und EXIT-Spielen für die verschiedensten Themengebiete und Altersstufen entwickelt. Mit „Breakout EDU“ existiert mittlerweile eine internationale Onlineplattform mit deren Unterstützung Lehrer*innen digitale ESCAPE-Räume ohne großen Aufwand in ihren Unterricht implementieren können [25]. Aber auch klassische Lehrmittelfirmen und Verlage haben das Potential dieser speziellen Art des Unterrichts erkannt und ebenfalls Lehrmaterial herausgebracht. Eine Vorreiterrolle hat hier im deutschsprachigen Raum der AUER-Verlag übernommen, der bereits Breakout-Spiele in Kombination aus Buchform und Onlinematerial für Lehrer*innen an Grund- und weiterführenden Schulen in verschiedenen Unterrichtsfächern wie z.B. Biologie, Mathematik, Deutsch, Geschichte und Religion zur Verfügung stellt.

2.4. Die Rolle der Motivation beim Einsatz von EXIT-Spielen

Bereits mehrfach wurde in vorangegangenen Kapiteln die Steigerung der Motivation von Schüler*innen durch das Spielen im Unterricht erwähnt, weshalb deren Rolle nachfolgend kurz betrachtet wird.

Motivation ist ein psychologischer Prozess, welcher den Beginn, die Durchführung, aber auch die Qualität des menschlichen Handels beeinflusst [32]. Weitergehend kann zwischen intrinsischer und extrinsischer Motivation unterschieden werden. Eine intrinsisch motivierte Handlung wird wegen der Handlung selbst ausgeführt, weil diese als interessant, spannend oder herausfordernd empfunden wird [20]. Extrinsisch motivierte Handlungen hingegen erfolgen nicht aus Freude an der Handlung, sondern lediglich um ein gewisses Ergebnis zu erzielen [33]. Das Forschungsinteresse zielt hierbei genau auf die intrinsische Motivation der Schüler*innen ab. Prax stellt fest, dass Spielen im Unterricht sich nachhaltig positiv auf das Interesse und die Motivation der Schüler*innen auswirkt [20]. Auch Hamari konnte empirisch belegen, dass Engagement in spielerischen Handlungen einen positiven Effekt auf das Erreichen von Lernzielen hat [34]. Wellenreuther schließt an diese Beobachtungen an und beschreibt Details wie den Titel und eine dem Spiel zu Grunde liegende Geschichte als maßgeblich für die Immersion und damit verbunden die Steigerung der intrinsischen Motivation [35]. Zusammenfassend kann festgehalten werden, dass die intrinsische Motivation durch das Spielen im Unterricht gesteigert

werden kann, wenn die Bedingungen der sozialen Eingebundenheit, des Kompetenzerlebens und die Erfüllung des Bedürfnisses nach Autonomie seitens der Schüler*innen gegeben ist [32]. Folglich wirkt sich der Spieleinsatz positiv auf das Lernpotential der Schüler*innen aus [7, 34].

3. Motivation, Fragestellung und Hypothesen der Arbeit

3.1. Motivation

Als wohl wesentlichsten Bestandteil der Motivation für diese Untersuchung stellt sich die Tatsache heraus, dass das Interesse und auch der Einsatz von ESCAPE-Räumen und auf diesem Konzept basierenden Spielen über die letzten Jahre stark zugenommen hat [24]. Lehrmittelhersteller haben diesen Trend für sich erkannt und erste speziell für den Unterricht entwickelte ESCAPE-Räume auf den Markt gebracht. Gleichzeitig besteht jedoch weiterhin ein großes Defizit in der empirischen Untersuchung der Effekte des Einsatzes dieser Spiele im Unterricht [11]. Dies wirft wiederum die Frage auf, ob eine Vermarktung und daraus folgende großflächige Verbreitung dieser Spielgattung empfehlenswert sind, solange deren tatsächlicher Einfluss auf Schüler*innen und deren Lernprozesse von Forschenden nicht ausreichend untersucht wurden.

Darüber hinaus lieferten aktuelle Studien eine Vielzahl an Daten, die zumindest den generellen Einsatz von Spielen im Unterricht als lernförderlich bestätigten [36, 37]. Besonders die in Skene´s Metaanalyse beobachtete Überlegenheit des spielerischen Lernens gegenüber klassischer Instruktion auf mathematische Kompetenzen hat zur Zielsetzung geführt, ähnlich positive Effekte mit der Durchführung des zu untersuchenden Spiels zu erreichen. Konkret haben bereits mehrere Untersuchungen einen positiven Effekt auch speziell bezogen auf ESCAPE-Räume oder ähnliche Spiele, deren inhaltlicher Fokus neben biologischem auch auf physikalischem Fachwissenserwerb lag und in zu dieser Untersuchung vergleichbaren Altersgruppen durchgeführt worden sind, ergeben [18, 23, 29]. Ein Großteil der aktuellen Forschungsprojekte hingegen hat vor allem motivationale und prozessbezogene Aspekte untersucht (s. Kapitel 1). Dahingehend bietet diese Untersuchung das Potential, eine bisher wenig untersuchte Rolle dieser Spielgattung, nämlich die Nutzung zum Erwerb, zur Festigung und zum Vertiefen von physikalischem Fachwissen, näher zu beleuchten.

Ferner ist anzumerken, dass das zu untersuchende Spiel bereits in einer vorangegangenen Arbeit vom Testleiter selbst entwickelt wurde [9]. Daher ist das Interesse groß mehr über die Wirkung des Spiels zu erfahren.

3.2. Forschungsfragen

Aus dem Forschungsinteresse resultieren zwei wesentliche Forschungsfragen (FF):

FF1: Inwieweit wird der Kompetenzzuwachs der Schüler*innen im kognitiven Bereich durch den Einsatz eines EXIT-Spiels im Vergleich zu konventionellen, vertiefenden Unterrichtsstunden beeinflusst?

FF2: Wie wird der Einsatz des EXIT-Spiels im Unterricht von den spielenden Schüler*innen wahrgenommen?

FF1 bezieht sich dabei auf den Kompetenzzuwachs der Schüler*innen, der mittels des entwickelten Leistungstests überprüft wurde und den Großteil dieser Arbeit ausmacht. Zusätzlich zum Kompetenzzuwachs der Schüler*innen wird anhand von FF2 die intrinsische Motivation der Schüler*innen durch den Spieleinsatz untersucht.

3.3. Hypothesen

Anhand der beiden Forschungsfragen lassen sich mehrere Hypothesen ableiten, die anhand der nachfolgend erhobenen Daten überprüft werden sollen. Die Hypothese H1 und dazugehörige Nullhypothese beziehen sich auf die die Forschungsfrage FF1 und werden über die Leistungstests überprüft. Die beiden Hypothesen H2 und H3 beziehen sich auf die Forschungsfrage FF2 und lassen sich anhand des Fragebogens überprüfen.

H1: Der Einsatz von EXIT-Spielen im Physikunterricht hat einen Einfluss auf den Kompetenzzuwachs der Schüler*innen.
H0: Es existiert kein Zusammenhang zwischen dem Kompetenzzuwachs der der Schüler*innen und dem Einsatz von EXIT-Spielen.

H2: Der Einsatz von EXIT-Spielen im Physikunterricht steigert die Motivation der Schüler*innen.
H3: Der Einsatz des EXIT-Spiels im Physikunterricht steigert das Interesse der Schüler*innen für die Kernphysik.

4. Untersuchungsdesign und Methoden

4.1. Forschungsmethodik

4.1.1. Quasiexperimentelles Untersuchungsdesign

Als Grundlage für die Untersuchung wurde ein quasiexperimentelles Pre-Posttest-Design gewählt. Der wesentliche Unterschied zu einer experimentellen Untersuchung besteht in der Auswahl und Randomisierung der Versuchsgruppen. Während in einer experimentellen Untersuchung die Versuchsgruppen vollständig randomisiert werden, um die Vergleichbarkeit der entsprechenden Experimental- und Kontrollgruppen zu erreichen, wird in einer quasiexperimentellen Untersuchung auf nicht randomisierte oder bereits bestehende Gruppen, wie beispielsweise Schulklassen, zurückgegriffen [38]. Die Zuordnung von Versuchspersonen zu den Untersuchungsgruppen wird in diesem Fall also nicht zufällig bestimmt, sondern unterliegt der möglichst simplen Realisierbarkeit der Untersuchung. Bierhoff und Rudinger sprechen in diesem Zusammenhang von „nichtäquivalenten Kontrollgruppenplänen“, da davon ausgegangen werden muss, dass sich die Experimental- und Kontrollgruppe von Beginn an in wichtigen Untersuchungsmerkmalen unterscheiden [39]. Die Möglichkeiten der Untersuchungsmethodik unterliegen aufgrund der Rahmenbedingungen von schulischen Institutionen gewissen Einschränkungen. Dazu zählen neben der vorgegebenen Klassenstruktur, die sich maßgeblich auf das Design auswirkt, noch andere Regularien, wie die Unterrichtszeit, curricularen Vorgaben und die Genehmigung der Durchführung an Schulen von Seiten der Schulleitung. Damit der Einfluss möglicher Störvariablen so gering wie möglich ausfällt, sollte laut Döring die Datenerhebung in einem hohen Maß parallelisiert durchgeführt werden [38]. Daher wurden für die Teilnahme insgesamt vier zehnte Klassen von zwei Schulen im hannoverschen Stadtgebiet ausgewählt. Genaueres zur Auswahl der beiden Schulen und auch der Klassen folgt im Kapitel 4.2.

Im Pre-Posttest-Design erfolgt die Veränderungsmessung über zwei Teilmessungen zu zwei unterschiedlichen Messzeitpunkten. So wird vor der Intervention (EXIT-Spiel) ein Pretest zur Messung des initialen Lernstandes durchgeführt. Nach erfolgter Intervention wird eine zweite Erhebung des Lernstands durchgeführt. Die Ermittlung von Lerneffekten durch das Spiel erfolgt durch Differenzbildung der beiden Testergebnisse und einem Vergleich der Experimental- und Kontrollgruppen. In Tabelle 4 ist dieser Prozess schematisch in der für diese Untersuchung gewählten Codierung dargestellt.

Tabelle 4: Schematische Darstellung zur Ermittlung der Lerneffekte aller Versuchsgruppen [38].

Schule	**Untersuchungsgruppe**	**Pretest**	**Posttest**	**Differenz**
St. Ursula-Schule	Experimentalgruppe	UVE	UNE	UDE = UNE – UVE
	Kontrollgruppe	UVK	UNK	UDK = UNK – UVK
			Nettolerneffekt	UNL = UDE – UDK
Bismarck-schule	Experimentalgruppe	BVE	BNE	BDE = BNE – BVE
	Kontrollgruppe	BVK	BNK	BDK = BNK – BVK
			Nettolerneffekt	BNL = BDE – BDK

Bemerkung: Zur Abkürzung der Versuchsgruppen wurde folgende Codierung genutzt: **B** Bismarckschule; **U** St. Ursula-Schule; **V** Pretest; **N** Posttest; **E** Experimentalgruppe; **K** Kontrollgruppe; **D** Differenz.

In der Literatur wurde der Einsatz von einfachen Differenzmaßen aus Pre- und Posttest-daten oftmals aufgrund mangelhafter Reliabilität kritisiert. Döring und Bortz sind jedoch der Meinung, dass gerade bei Untersuchungen, in denen es organisatorisch nicht anders möglich ist, wie beispielsweise in der Schule, der Einsatz durchaus angebracht und aussagekräftig ist [38]. Zur Vermeidung der Reliabilitätsproblematik wird das entwickelte Messinstrument in Kapitel 5.3 mithilfe verschiedener Gütekriterien überprüft. Weiterführend müssen Effekte, wie die Regression zur Mitte bei der Anwendung dieses Versuchsdesigns beachtet werden. Als Regression zur Mitte wird ein Effekt bezeichnet, bei dem Posttest-Werte zur Mitte der Punkteverteilung tendieren, obwohl im Pretest Extremwerte auf beiden Seiten, also besonders hohe und niedrige Gesamtpunktzahlen, erreicht worden sind [38]. Der Einfluss lässt sich minimieren, wenn wie in dieser Untersuchung verschiedene Klassen von verschiedenen Schulen ausgewählt werden [39]. Zur Gewährleistung der Vergleichbarkeit der Pre- und Posttest-Daten, waren die Frageitems und Antwortoptionen beider Tests identisch, lediglich die Reihenfolge der Antwortoptionen wurde randomisiert. Jedoch können zwischenzeitliche Lerneffekte aufgrund des Pretests, sogenannte Sequenzeffekte (z.B. Klassenarbeiten), die die Posttest-Ergebnisse zu höheren Punktzahlen verfälschen, nicht ausgeschlossen werden [38].

4.1.2. Datenerhebung

Die Datenerhebung war für einen vierwöchigen Zeitraum zum Ende des Schuljahres 2021/22 vorgesehen. Damit die Schüler*innen ausreichend Vorwissen, das im Spiel benötigt wird, aufbauen konnten, wurde der Erhebungszeitraum knapp vor den Sommerferien gewählt. Außerdem steht zum Schuljahresende meist ausreichend Unterrichtszeit zur

Verfügung, da Klassenarbeiten bereits geschrieben sind und auch sonstiger Unterrichtsausfall (z.B. durch Feiertage) in den Monaten Juni und Juli selten ist. In diesem knapp kalkulierten Zeitraum sollten die vier Versuchsgruppen an der Untersuchung teilnehmen. An jeder Schule war eine Experimentalgruppe, die das EXIT-Spiel im Unterricht einsetzte, und eine Kontrollgruppe, die mit dem regulären Unterricht fortfuhr, vorgesehen. Hierbei sollten alle vier Versuchsgruppen möglichst zur selben Zeit, oder zumindest in derselben Woche die drei Stufen der Datenerhebung durchlaufen. Die Pre- und Posttest-Stufe nahm jeweils eine Unterrichtsdoppelstunde pro Klasse und damit verbunden je eine Woche des Zeitplans ein. Die Spielphase hingegen hatte eine Länge von zwei Unterrichtsdoppelstunden, benötigte damit auch zwei Wochen. Somit ergibt sich zum Ende eine Gesamtzeitspanne von vier Wochen für den Prozess der Datenerhebung. Der zeitliche Ablauf wird schematisch in Abbildung 2 dargestellt. Aus dieser wird die zeitliche und auch räumliche Trennung der Phasen von Leistungstests und Spiel ersichtlich. Diese Trennung ist für die Schüler*innen erforderlich, um eine Separation von Lern- und Leistungssituation zu ermöglichen. Leisen sieht dies als Grundlage einer aussagekräftigen Diagnostik an, da Lern- und Leistungssituationen verschiedene psychologische Bedingungen erfordern. Diese müssen den Schüler*innen durch eine deutliche Trennung der Phasen kenntlich gemacht werden, damit diese adäquat auf die jeweilige Situation reagieren können [40].

Der Erhebungszeitraum verlängerte sich aufgrund terminlicher Differenzen von vier auf sechs Wochen (01.06. – 06.07.2022), sodass auch die angestrebte parallelisierte Durchführung an beiden Schulen nur bedingt möglich war. Dies hatte zur Folge, dass der Pretest an der Bismarckschule bereits zwei Wochen vor dem Beginn der Datenerhebung an der St. Ursula-Schule durchgeführt werden musste. Ferner verdoppelte sich an der Bismarckschule der zeitliche Abstand zwischen Pre- und Posttest von zwei auf vier Wochen. Auch auf den Einfluss dieser unvorhergesehenen Änderung wird im Kapitel 8 weiterführend eingegangen. Die verbleibenden zwei Phasen der Datenerhebung, Spiel und Posttest, konnten dann jedoch wieder an beiden Schulen parallel ablaufen, sodass die Datenerhebung dann pünktlich in der letzten vollständigen Unterrichtswoche vor den Sommerferien endete.

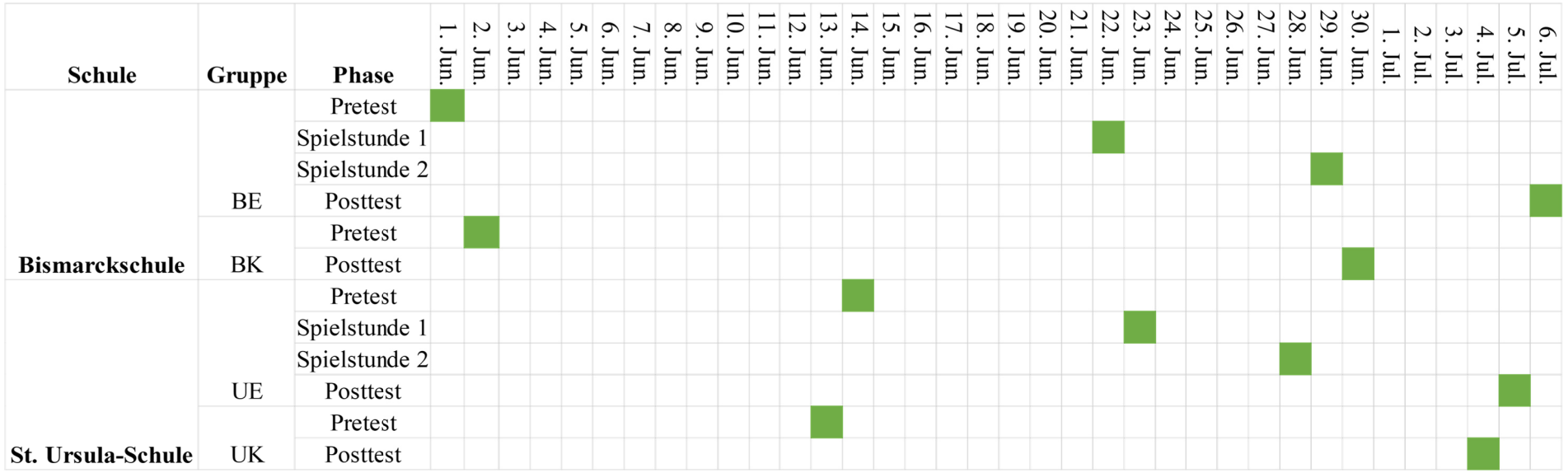


Abbildung 2: Gantt-Diagramm des zeitlichen Ablaufs der Datenerhebung.

Bemerkung: Zur Abkürzung der Versuchsgruppen wurde folgende Codierung genutzt: **B** Bismarckschule; **U** St. Ursula-Schule; **E** Experimentalgruppe; **K** Kontrollgruppe.

4.2. Auswahl der Untersuchungsgruppen

Eine besondere Herausforderung in der Vorbereitung dieser Untersuchung stellte die Auswahl der Versuchsgruppen dar. Hierbei ist auf persönliche Kontakte zu entsprechenden Fachlehrkräften zurückgegriffen worden. Die den Auswahlprozess betreffenden Faktoren sollen nachfolgend genauer aufgeschlüsselt werden.

Zur Gewährleistung der Vergleichbarkeit aller Versuchsgruppen mussten gewisse Bedingungen erfüllt werden. Hierzu zählten in erster Linie, (1) dass die Schulform aller Klassen identisch war, (2) dass die Jahrgangsstufe aller Klassen identisch war und (3) dass die Experimental- und die dazugehörige Kontrollgruppe von derselben Lehrkraft unterrichtet wurden. Die gestellten Bedingungen reduzierten die Zahl geeigneter Schulen. Wie bereits eingangs erwähnt, sollten an der Untersuchung zwei Schulen mit jeweils zwei Klassen teilnehmen. Eine weitere Bedingung war die thematische Auseinandersetzung mit dem Themenkomplex „Kernphysik“. Dieser wird im Physikunterricht der Sekundarstufe I im Doppeljahrgang 9/10, sowie im Physikunterricht der gymnasialen Oberstufe (13. Jahrgang) abgehandelt [31, 41]. Durch diese curriculare Einschränkung beschränkte sich die Suche nach potentiellen Teilnehmenden auf weiterführende Schulen (Gymnasien, Gesamtschulen, Oberschulen, etc.) und die drei Jahrgänge. Der Erhebungszeitraum vor Beginn der Sommerferien führte dazu, dass die Physikkurse der gymnasialen Oberstufe nicht an der Erhebung teilnehmen konnten, da die Abiturprüfungen in diesen Zeitraum fielen. Ferner wurde das Themenfeld „Atom- und Kernphysik“ aufgrund der Nachwirkungen der Corona-Pandemie aus dem Anforderungsprofil für die Abiturprüfungen im Fach Physik gestrichen [42]. Eine Durchführung der Untersuchung wäre somit auch bei vorhandenen zeitlichen Kapazitäten nicht ratsam gewesen.

Diese Entscheidung des niedersächsischen Kultusministeriums sorgte für eine Festlegung auf den Doppeljahrgang 9/10. Da das EXIT-Spiel ursprünglich für die gymnasiale Oberstufe konzipiert worden ist, mussten vor dem Einsatz einige Anpassungen vorgenommen werden, welche in Kapitel 5.1 ausführlich beschrieben werden.

Zur Festlegung auf einen speziellen Untersuchungsjahrgang musste beachtet werden, dass das niedersächsische Kultusministerium für das Unterrichtsfach Physik in der Sekundarstufe I nur in sogenannte Doppeljahrgänge differenziert [31]. Es obliegt den jeweiligen Fachkonferenzen der Schulen festzulegen, welche Inhalte und Themenkomplexe in welchem Schuljahr des Doppeljahrganges behandelt werden. So ist zu beobachten, dass die Kernphysik in einigen Schulen bereits im Jahrgang 9 und an anderen erst im Jahrgang

10 unterrichtet wird. Diese Beobachtung ist für die nachfolgend beschriebene Auswahl der Schulen von großer Bedeutung.

4.2.1. Schulauswahl

Die in der hannoverschen Südstadt ansässige St. Ursula-Schule ist ein Gymnasium in privater Trägerschaft der Stiftung Katholische Schule in der Diözese Hildesheim. Die Schüler*innenschaft stellt sich als motivierte, engagierte und vor allem leistungsstarke Gemeinschaft dar, in der es nur sehr selten zu Komplikationen und Konflikten kommt. Ausgewählt wurde diese Schule primär aufgrund persönlicher Kontakte zu den dort unterrichtenden Physiklehrer*innen. Neben der damit verbundenen niedrigschwelligen Kontaktaufnahme und Organisation, war ein weiterer Punkt ausschlaggebend. Der Faktor Lehrkraft und deren erteilter Unterricht wirkt sich stark auf die zu erhebenden Daten aus. Aus diesem Grund sollten, wenn möglich, die Kontroll- und Experimentalgruppe an der jeweiligen Schule von derselben Lehrkraft unterrichtet werden (s. Kapitel 4.2). Dieses Kriterium wurde von einem Lehrer erfüllt, sodass sich an der St. Ursula-Schule zwei 10. Klassen für die Datenerhebung finden ließen.

Zur Sicherung der Vergleichbarkeit wurde eine äquivalente Schule mit zwei Klassen des 10. Jahrgangs gesucht. In der Bismarckschule, einem der ältesten Gymnasien in Hannover, wurde eine Schule gefunden, die diese Bedingungen erfüllt und an dem sich neben einem betreuenden Physiklehrer, auch die Schulleitung für das Projekt begeistern ließ. Auch in diesem Fall konnte auf vorherige Kontakte zurückgegriffen werden. Die Schüler*innen ähneln denen der St. Ursula-Schule hinsichtlich Sozialverhalten und Leistungsspektrum. Daher eignete sich diese Schule für einen Vergleich in besonderem Maße.

4.2.2. Versuchsgruppenauswahl

Die Einteilung der Versuchsgruppen wurde in Absprache mit den Fachlehrkräften vorgenommen. Hierbei sollten die Gruppen möglichst randomisiert zugeteilt werden, um die Ergebnisse nicht durch das Zuordnen der leistungsstarken/-schwachen Klassen zu verfälschen. Letztendlich erfolgte die Zuordnung zu den Versuchsgruppen jedoch an beiden Schulen aus terminlichen Gründen, da die Datenerhebung ansonsten nicht möglich gewesen wäre.

So wurde an der St. Ursula-Schule die von der Fachlehrkraft als leistungsstärker eingeschätzte Klasse die Experimentalgruppe und an der Bismarckschule hingegen eine Lerngruppe, die der Physiklehrer als eher leistungsschwach und chaotisch beschrieben hatte. Die Kontrollgruppen setzten sich demnach genau entgegengesetzt zusammen. Hier war

die leistungsstärkere Klasse an der Bismarckschule zu finden, wohingegen die leistungsschwächere Klasse an der St. Ursula Schule die Kontrollgruppe stellte.

4.3. Demographie der Untersuchungsgruppen

Die demographischen Daten wurden während des Pretests erhoben. Aufgrund der anhaltenden Corona-Pandemie fehlten einige Schüler*innen beim Pretest, sodass die nachfolgend erhobenen Daten nicht die vollständigen Klassen abbilden. Weiterführend muss berücksichtigt werden, dass im Laufe der Intervention bis zur Datenerhebung einige Schüler*innen genesen, andere dafür erkrankt sind, sodass sich durchweg eine gewisse Fluktuation an Schüler*innen über den gesamten sechswöchigen Erhebungszeitraum einstellte. Da es sich bei den Schulnoten um selbst berichtete Daten handelt, ist eine kritische Betrachtung nötig. Insgesamt nahmen an dieser Untersuchung 96 Schüler*innen (M_{Alter} = 15,75 Jahre, SD_{Alter} = 0,54 Jahre, Altersspanne von 15 bis 17 Jahre, 48% weiblich, 2% divers) teil.

4.3.1. St. Ursula-Schule

Am Vortest der Experimentalgruppe an der St. Ursula-Schule nahmen 26 Schüler*innen teil. Zwei Datensätze konnten jedoch aufgrund eines sich doppelnden Teilnehmenden-Codes nicht berücksichtigt werden. Die Stichprobe bestand somit aus 24 Schüler*innen (M_{Alter} = 15,79 Jahre, SD_{Alter} = 0,58 Jahre, Altersspanne von 15 bis 17 Jahre, 54% weiblich, 8% divers). Die Physiknoten (selbst berichtet) verteilten sich wie folgt: Note 1 (12%), Note 2 (21%), Note 3 (38%), Note 4 (29%). Abschließend wurde noch die Vorerfahrung der Schüler*innen im Umgang mit ESCAPE-Räumen und EXIT-Spielen erhoben. Hierbei gaben 38% der Schüler*innen an, noch nie ein solches Spiel gespielt zu haben. Ebenfalls 38% hatten bereits ein solches Spiel gespielt, 20% haben Erfahrung mit zwei bis fünf und 2% mit mehr als fünf dieser Spiele.

Die Kontrollgruppe der St. Ursula-Schule bestand aus 23 Schüler*innen (M_{Alter} = 15,83 Jahre, SD_{Alter} = 0,49 Jahre, Altersspanne von 15 bis 17 Jahre, 46% weiblich). Die Schulnoten setzten sich hier folgendermaßen zusammen: Note 1 (23%), Note 2 (19%), Note 3 (27%), Note 4 (27%), Note 6 (4%). Bezüglich der Vorerfahrungen mit ESCAPE-Räumen gaben 58% der Schüler*innen an, diese noch nie besucht oder ähnliches gespielt zu haben, 23% hatten bereits einmal, 15% zwischen zwei und fünf Mal und 4% mehr als fünf Mal Kontakt mit dieser Art von Live-Action-Unterhaltung.

4.3.2. Bismarckschule

Die Experimentalgruppe der Bismarckschule bestand aus 26 Schüler*innen (M_{Alter} = 15,73 Jahre, SD_{Alter} = 0,60 Jahre, Altersspanne von 15 bis 17 Jahre, 50% weiblich). Die Zeugnisnoten verteilten sich in dieser Klasse auf 12% der Note 1, 34% der Note 2, 27% der Note 3 und 27% der Note 4. Analog zu den anderen beiden Versuchsgruppen hatten auch hier der Großteil der Schüler*innen noch nie Kontakt mit EXIT-Spielen oder Ähnlichem (62%). Auffällig ist hier jedoch, dass an zweiter Stelle der Häufigkeit bereits die Option zwischen zwei und fünf maligen Spielens gewählt wurde (27%). Lediglich 8% hatten nur einmal und 4% mehr als fünf Mal Kontakt mit ESCAPE-Räumen oder EXIT-Spielen.

Die Kontrollgruppe der Bismarckschule, bestand aus 23 Schüler*innen (M_{Alter} = 15,65 Jahre, SD_{Alter} = 0,48 Jahre, Altersspanne von 15 bis 16 Jahre, 43% weiblich). Die Physiknoten aus dem vorangegangenen Schulhalbjahr verteilen sich mit 43% auf die Note 2, 22% auf die Note 3 und 35% auf die Note 4. In dieser Gruppe hatten auch die wenigsten Schüler*innen Erfahrungen mit ESCAPE-Räumen. 70% gaben an noch nie in solch einem Spiel teilgenommen zu haben, 13% hatten dies bereits einmal getan, weitere 13% bereits mehr als fünf Mal und lediglich 4% gaben an zwischen zwei und fünf Mal in einem ESCAPE-Raum gewesen zu sein oder ein EXIT-Spiel gespielt zu haben.

5. Entwicklung der Erhebungsinstrumente

5.1. Anpassung des EXIT-Spiels

Ursprünglich wurde das Spiel „Escape – Gefangen bei der Wismut“ für den Physikunterricht in der gymnasialen Oberstufe konzipiert [9]. Die Inhalte des Spiels sind an die inhaltsbezogenen Kompetenzen des niedersächsischen Kerncurriculums angeglichen, um so die Einbindung in den Unterricht legitimieren zu können [41]. In dieser Untersuchung nehmen jedoch Schüler*innen des 10. Jahrganges aus bereits in Kapitel 4.2 beschriebenen Gründen teil. Damit die Rätsel des Spiels für die niedrigere Jahrgangsstufe der spielenden Schüler*innen auch weiterhin zu bewältigen sind und das Spiel damit seinem intendierten Zweck weiterhin nachkommen kann, müssen einige Anpassungen vorgenommen werden. Die größte Anpassung war die Kürzung des Spiels von 15 auf zehn Rätsel. Diese Kürzung hatte vorwiegend inhaltliche Gründe, da die in den Kerncurricula beschriebenen inhaltsbezogenen Kompetenzen für die Sekundarstufe II umfangreicher ausfallen als die für den Doppeljahrgang 9/10 [31, 41]. Daher mussten Rätsel, die beispielsweise die Strahlungswechselwirkung mit Materie, den Uranbrennstoffkreislauf oder den Aufbau und die Funktionsweise moderner Kernkraftwerke thematisieren, aus dem Spiel ausgeschlossen werden. Weiterhin zeigte sich in den ersten Erprobungen des Spiels die Notwendigkeit einer größeren Bearbeitungszeit. Damit das Spiel im regulären Unterricht einsetzbar bleibt, musste die Bearbeitungszeit entweder durch eine geringere Komplexität der einzelnen Rätsel, oder eine Reduzierung der Rätselanzahl verringert werden. In Kombination mit den vorangegangenen Überlegungen ist die Rätselanzahl reduziert worden. Ob sich ein Erfolg dieser Maßnahme, in Form der erhofften Verkürzung der Bearbeitungszeit auf maximal zwei Unterrichtsdoppelstunden, also 180 Minuten, eingestellt hat, wird in einem nachfolgenden Kapitel dieser Arbeit thematisiert.

Neben der inhaltlichen Reduktion des Spiels wurde auf Anregung der Lehrkräfte eine zusätzliche Anleitung erstellt, die die notwendigen Anpassungen von der Version für die gymnasiale Oberstufe auf die Spielvariante der Sekundarstufe I beschreibt (s. Anhang). Hierbei ist es gelungen, den Anpassungsaufwand für die Lehrkräfte so gering wie möglich zu halten. Von den mehr als 200 Spielkarten müssen lediglich 13 ausgetauscht und 39 vollständig aus dem Spiel entfernt werden. Drei weitere Spielgegenstände müssen entfernt und durch vier abgewandelte ersetzt werden. Das benötigte Tauschmaterial wird in einem zusätzlichen Umschlag dem bestehenden Spiel ergänzt und stellt den einsetzenden Lehrkräften somit ein Komplettpaket zur Verfügung. Interessierten Lehrkräften stehen

daher mit einem geringen organisatorischen Aufwand sowohl eine Spielvariante für die gymnasiale Oberstufe als auch eine für die Sekundarstufe I zur Verfügung.
Als direkte Folge der Kürzung des Spiels ergab sich die Notwendigkeit der Anpassung der Rahmenerzählung und fortlaufenden Geschichte, den Spielenden auf Lösungs- und Rätselkarten mitgeteilt wird. So musste an der Stelle jedes ausgelassenen Rätsels die Geschichte geringfügig angepasst werden, um inhaltliche Brüche und Unstimmigkeiten, die das „Flow-Erleben“ beeinflussen würden, zu verhindern. Durch geschickte Verknüpfungen und den günstigen Zufall, dass sich nahezu alle Rätsel, die aus dem Spiel ausgeschlossen wurden, an dessen Ende befanden, war es möglich, die Spielgeschichte auf eine sinnvolle Art und Weise zu kürzen. So ist lediglich ein kurzer narrativer Abschnitt vor der erfolgreichen Flucht aus der Mine nicht mehr im Spiel enthalten. Die hierfür neugestalteten Spielkarten und das zusätzliche Material sind ebenfalls dem digitalen Anhang zu entnehmen.

5.2. Konstruktion des Messverfahrens

Nachdem das Spiel an die Bedingungen der Untersuchungsgruppen angepasst worden ist, galt es, im Anschluss ein Messinstrument zu konstruieren, welches in der Lage ist die im vorangegangenen Unterricht angeeigneten inhaltsbezogenen Kompetenzen der Schüler*innen und auch die durch das Spiel hinzugekommenen oder verstärkten Kompetenzen aussagekräftig zu erheben. In diesem Fall wird auch von summativer Diagnostik gesprochen [43]. Hierbei soll das Messinstrument ausschließlich den Bereich der sogenannten kognitiven Leistungen erfassen. Laut Häußler gehören in diesen Bereich der kognitiven Leistungen das Wissen von Einzelheiten und Benennungen, das Wissen über Begriffe und Theorien, das Verstehen von Zusammenhängen, höhere kognitive Leistungen und das Bewerten [44]. Das entwickelte Messinstrument hingegen beschränkt sich nur auf die ersten drei genannten Kategorien kognitiver Leistung.
Bei der Entwicklung des Leistungstest wurde auf ein bewährtes Verfahren in vier Schritten nach Worbach, Drechsel und Carstensen zurückgegriffen [45]. Diese Vorgehensweise sieht zu Beginn eine Präzisierung des zu messenden Konstrukts, also die Festlegung von Lerninhalten vor. Anschließend folgt die Ausgestaltung der Testitems mit der Entwicklung zugehöriger Antwortoptionen und deren Bewertungsmaßstab im nachfolgenden Arbeitsschritt. Zum Abschluss erfolgt eine Überprüfung des Messinstruments anhand verschiedener Gütekriterien [45]. Sacher schlägt eine kleinschrittigere Vorgehensweise vor,

deren wesentliche Konstruktionsschritte hingegen deckungsgleich mit denen von Worbach, Drechsel und Carstensen ausfallen und daher nachfolgend ebenfalls Berücksichtigung finden [46].

5.2.1. Präzisierung des Konstrukts

Zu Beginn der Entwicklung stand also die Festlegung des zu messenden Gegenstands, in diesem Falle dem Fachwissen zur Kernphysik, welches für die erfolgreiche Durchführung des Spiels benötigt werden würden. Hierzu geben die inhaltsbezogenen Kompetenzen des niedersächsischen Kerncurriculums wesentliche Anhaltspunkte [31]. Zusammen mit den thematischen Schwerpunkten des Spiels, die bereits mit dem Kerncurriculum abgestimmt worden sind, ergeben sich sieben Themenbereiche, die das Messinstrument abdecken soll [9]:

Tabelle 5: Kernphysikalische Themenbereiche mit dazugehörigen Skalenabkürzungen.

Themenbereich	**Skalenabkürzung**
Der Aufbau und die Bestandteile des Atoms	FA
Die Arten ionisierender Strahlung und deren Eigenschaften	FB
Die Detektion ionisierender Strahlung	FC
Die Karlsruher Nuklidkarte [Vgl. 47]	FD
Der radioaktive Zerfall	FE
Die natürlichen Zerfallsreihen	FF
Die neutroneninduzierte Kernspaltung	FG

Mit dem Abschluss der Unterrichtseinheit zur Kernphysik im Doppeljahrgang 9/10 sollte ein gewisses Maß inhaltsbezogener Kompetenzen erreicht worden sein, damit diese durch den Spieleinsatz wiederholt, vertieft und gefestigt werden können. Zur angemessenen Beurteilung der Schülerleistung muss vor der Entwicklung des Erhebungsinstruments die Mindestanforderung an die Schüler*innen definiert werden, die als vorausgesetzt angenommen werden kann [Vgl. 48]. Anhand der vorher festgelegten Themenbereiche lassen sich folgende Lernziele für die kognitiven Leistungen formulieren. Bei der Darstellung sind Anforderungen, die über die im Kerncurriculum geforderten Kompetenzen hinausgehen, *kursiv* dargestellt.

Die Schüler*innen sind im Einzelnen in der Lage:

- die Bestandteile des Atoms zu benennen und dessen Aufbau zu beschreiben,

- die drei wesentlichen Arten ionisierender Strahlung (α-, β- und γ-Strahlung) zu benennen, deren Eigenschaften zu beschreiben und anhand dieser das Gefährdungspotential ionisierender Strahlung zu beurteilen,
- den Aufbau und die Funktionsweise moderner Geräte zur Strahlungsdetektion (Geiger-Müller Zählrohr, *Halbleiterdetektor, etc.*) unter Zuhilfenahme von Fachbegriffen zu erklären,
- *die Karlsruher Nuklidkarte sicher anzuwenden und die darin enthaltenen Informationen zu deuten* [47],
- den stochastischen Charakter des radioaktiven Zerfalls, sowie die damit verbundene Bedeutung der Halbwertszeit unter Zuhilfenahme von Fachbegriffen zu erläutern und mittels der mathematischen Gesetzmäßigkeiten Berechnungen durchzuführen,
- *die vier natürlichen Zerfallsreihen, sowie deren Start- und Endnuklide zu benennen und über das 4n-Schema den Zerfallsreihen zugehörige Radionuklide zu bestimmen,*
- den Prozess der neutroneninduzierten Kernspaltung zu erklären, sowie die militärische und zivile Nutzung anhand verschiedener Kriterien zu bewerten und begründet Stellung zu beziehen.

5.2.2. Formulierung der Items

Im Anschluss an die Formulierung der abzuprüfenden Lernziele mussten die einzelnen Aufgaben und Fragestellungen entwickelt werden. Dabei galt es, zu Beginn ein entsprechendes Item-Format festzulegen. Die Wahl ist hier auf ein gebundenes Format in Form von Single-Choice-Fragen mit jeweils fünf Antwortoptionen gefallen [48]. Dieses Item-Format wird aufgrund seines hohen Grades an Präspezifikation als gebunden betitelt, da alle wesentlichen Kriterien der Items (Lernziele, Fragestellungen, Antwortoptionen und Bewertungskriterien) vor der Durchführung des Leistungstests von der Versuchsleitung festgelegt worden sind [45]. Häußler bestätigt diese Wahl durch seine Feststellung, dass gerade Ankreuztests besonders gut geeignet sind, um die für diese Untersuchung relevanten Kategorien kognitiver Leistung (Wissen von Einzelheiten und Benennungen, das Wissen über Begriffe und Theorien, das Verstehen von Zusammenhängen) zu erfassen [44]. In Anlehnung an diese drei Kategorien wurden neben Single-Choice-Items und rein textbasierten Fragen auch Items mit Bildmaterial, zum Beispiel Ausschnitten aus der Nuklidkarten, oder Reaktionsgleichungen und Formeln versehen [47]. Infolgedessen war es

möglich, neben dem reinen Wissen und Benennen von Einzelheiten auch das Verständnis und die Anwendung von Theorien und Zusammenhängen mit diesem Instrument zu messen. Ein weiterer Vorteil der vorab festgelegten Anforderungen stellt die verbesserte Kommunikation dar [45]. Im generellen Schulalltag können sich die Schüler*innen auf Tests besser vorbereiten, wenn bereits im Vorfeld klar kommuniziert wird, welche Anforderungen erfüllt werden müssen. Für diese Untersuchung bot sich mit der Vorausplanung die Gelegenheit, frühzeitig mit den beiden unterrichtenden Fachlehrkräften in den Austausch zu treten. So konnten die Unterrichtsinhalte gemeinschaftlich auf die beschriebenen Ziele hin angepasst werden und gegebenenfalls an bestimmten Stellen nachgesteuert werden, sodass die Schüler*innen vorbereitet in die Leistungstests starten konnten.
Zur Erhöhung der Qualität der Messergebnisse wurde der Leistungstest mittels des Onlinetools "LimeSurvey" erstellt. Dies ermöglichte eine vollständig elektronische Datenerhebung und anschließende Auswertung, die neben vielen weiteren Kriterien von Bühner als maßgeblich für die Testqualität beschrieben wird [49]. Das beschriebene Erhebungsinstrument wurde zu beiden Messzeitpunkten (Pre- und Posttest) eingesetzt. Die Aufgaben, sowie deren Reihenfolge, wurden nicht verändert, um so vor allem die Vergleichbarkeit der Pre- und Posttests zu gewährleisten. Bei alternierenden Aufgaben im Posttest hätte nicht sichergestellt werden können, dass die Items die gleichen Kompetenzen abprüfen und dass die Verständlichkeit analog zum Pretest ist. Die bereits beschriebenen Themenbereiche dienten bei der Gestaltung des Leistungstest als Oberbegriffe zur Konstruktion von Itemskalen. Hieraus resultierten sieben Skalen, die jeweils aus vier bis fünf Items bestanden, welche für die Leistungsfeststellung genutzt wurden. Zusätzlich wurde eine weitere Fragengruppe zur Erhebung der demographischen Daten in den Pretest integriert, die im Posttest dann jedoch nicht ein weiteres Mal implementiert worden ist. Insgesamt beinhaltete der Test 30 inhaltliche Items zur Kernphysik und vier Fragen zur Erhebung der demographischen Zusammensetzung der Versuchsgruppen. Eine vollständige Auflistung aller Items des Leistungstest findet sich im Anhang dieser Arbeit.

5.2.3. Kategorisierung möglicher Antworten

Zum Abschluss der Konstruktion des Messinstruments galt es, zu jedem Item passende Antwortoptionen zu entwerfen und diese zu kategorisieren. Zur Vermeidung der Verzerrung von Daten durch das Raten einer Antwort wurde neben den fünf Antwortoptionen, aus denen es zu wählen galt, auch eine sechste Auswahlmöglichkeit „keine Antwort“ in jedes Item eingebaut. Die teilnehmenden Schüler*innen wurden ausdrücklich instruiert, sofern sie sich bei der Lösung einer Aufgabe unsicher waren, oder gar keine Antwort auf

eine Frage finden konnten, nicht zu raten, sondern die Option „keine Antwort“ auszuwählen. Die verbleibenden fünf Antwortoptionen bestanden für alle 30 Items stets aus nur einer richtigen Option (kodiert mit dem Wert „1“) und vier falschen Antwortoptionen (kodiert mit dem Wert „0“), den sogenannten „Distraktoren“ [44]. Mit dieser Bewertungsrichtlinie konnten die Schüler*innen somit maximal 30 mögliche Punkte erreichen. Aus den Datensätzen ist neben der reinen Bewertung richtig (1), oder falsch (0) zusätzlich durch eine fehlende numerische Angabe zu erkennen ob die Option „keine Antwort“ ausgewählt worden ist. Zusätzliche Differenzierungen in nur teilweise richtige Antworten mussten aufgrund des gewählten Single-Choice-Designs ebenfalls nicht entwickelt werden, was die anschließende Auswertung der Ergebnisse zusätzlich erleichterte.

Die Gestaltung der Distraktoren stellte eine besondere Herausforderung dar. Diese durften weder zu leicht noch zu schwer als falsche Antwort zu identifizieren sein. Daher wurden bei der Gestaltung dieser oftmals auf Schülervorstellungen zur Kernphysik zurückgegriffen. Das bietet den Vorteil, direkt aus den Daten zu ersehen, ob die Schüler*innen noch an diesen Vorstellungen festhalten, oder diese aktiv im Unterricht zu den physikalischen Vorstellungen umgedeutet werden konnten. So besteht eine Vielzahl von falschen Antwortoptionen aus weit verbreiteten Schülervorstellungen, wie beispielsweise dem Begriff der „radioaktiven Strahlung“, bei der die Schüler*innen die Eigenschaft des Aussendens von Strahlung (Radioaktivität) auf die ausgesendete ionisierende Strahlung übertragen [50]. Ein weiteres Beispiel sind nicht sachgemäße Vorstellungen zur Rolle der Halbwertszeit bei der exponentiellen Abnahme der Aktivität. Ein klassisches Beispiel einer Schülervorstellung zur Halbwertszeit wäre, dass sämtliches radioaktive Material nach bereits zwei Halbwertszeiten vollständig zerfallen ist.

Die verwendeten Antwortoptionen sind ebenfalls zusammen mit den Items vollständig im Anhang dargestellt. Aus dieser Auflistung wird auch das Bewertungsmodell ersichtlich, die jeweils richtige Antwortoption ist anschaulich markiert.

5.2.4. Überprüfung des Messmodells

Nach der Fertigstellung des Messinstruments gilt es, dieses anhand verschiedener Gütekriterien zu überprüfen. Für den gesamten Leistungstest erfolgt dies im folgenden Kapitel anhand von drei wesentlichen Hauptgütekriterien [51]. In diesem Kapitel werden die einzelnen Items auf ihre Schwierigkeit und Trennschärfe hin untersucht. Hierbei ergibt sich die Schwierigkeit der einzelnen Items aus dem Prozentwert der richtigen Lösungen [45, 52]. Es muss also beachtet werden, dass ein hoher Schwierigkeitsindex für ein leichtes Item spricht, da die Mehrheit der Teilnehmenden das Item beantworten konnte [38]. Die

Trennschärfe hingegen wird durch die Stärke der Korrelation von Item-Antworten und Gesamttestwert ausgedrückt [45, 52]. Schüler*innen die im Gesamtergebnis eine hohe Punktzahl erreicht haben, werden also ein trennscharfes Item wahrscheinlich richtig beantworten und umgekehrt [38]. Die Berechnung der Daten zur Häufigkeitsverteilung und Korrelation erfolgten mit der Statistiksoftware SPSS und sind in Tabelle 6 dargestellt [53, 54, 55, 56]. Diese und alle weiteren Korrelationen wurden nach dem Modell von Pearson über folgende Formel bestimmt:

$$r = \frac{cov_{xy}}{S_x \cdot S_y} = \frac{\frac{1}{N-1}\sum(x-\bar{x})\cdot(y-\bar{y})}{\sqrt{\frac{1}{N-1}\sum(x-\bar{x})}\cdot\sqrt{\frac{1}{N-1}\sum(y-\bar{y})}}$$

Formel 1: Korrelationskoeffizient nach Pearson [Vgl. 53, 54].

Bemerkung: cov_{xy} = Covarianz der Variablen x und y; S_x = Standardabweichung der Variable x; S_y = Standardabweichung der Variable y; $\bar{x}$ = Mittelwert der Variable x; $\bar{y}$ = Mittelwert der Variable y.

Für diese und alle folgenden Berechnungen wurden die Datensätze von insgesamt zwei Schüler*innen ausgeschlossen, da diese im Pretest der Experimentalgruppe an der St. Ursula-Schule einen exakt identischen Teilnehmenden-Code angegeben hatten und somit keine Differenzierung zwischen den Datensätzen möglich gewesen wäre. Aus den in Tabelle 6 dargestellten Werten wird ersichtlich, dass vor allem die Fragen zu den natürlichen Zerfallsreihen (FF1-FF4) für die Schüler*innen am schwierigsten zu beantworten waren. Auch die Frage bezüglich des primordialen Kalium-40 Nuklids (FD3) wurde nur von den wenigsten richtig beantwortet. Besonders einfache Aufgaben stellten dagegen die Fragen zu den Kernbestandteilen (FA1), der Sinneswahrnehmung von Strahlung (FC1) und zur Totzeit (FC4) dar. Damit aussagekräftige Ergebnisse getroffen werden können, sollte eine möglichst gleichmäßige Schwierigkeitsverteilung über alle Testitems vorherrschen. Besonders Extremwerte sind unerwünscht. Indizes zwischen $|.20| \leq p_{it} \leq |.80|$ sind wünschenswert, da individuelle Unterschiede besser erkannt werden können [38]. Bezüglich der Schwierigkeit des Messinstruments lässt sich also resümierend sagen, dass der Großteil der Items den Erwartungen entsprechend eine adäquate Schwierigkeitsverteilung aufweist. Jedoch existiert ein Überhang besonders schwieriger Items, die sich zusätzlich in einem ungünstigen Bereich des Indexes von $p_{it} \leq |.20|$ sammeln.

Tabelle 6: Schwierigkeit und Trennschärfe der Items des Leistungstests.

Item	Schwierigkeitsindex p_{it}	Trennschärfe r_{it}	Cronbachs-Alpha, wenn Item weggelassen
FA1	.79	.33	.76
FA2	.35	.36	.75
FA3	.43	.37	.75
FA4	.44	.49	.74
FB1	.35	.33	.75
FB2	.46	.35	.75
FB3	.54	.33	.75
FB4	.64	.34	.75
FC1	.82*	.31	.76
FC2	.23	.35	.75
FC3	.17*	.25*	.76
FC4	.75	.45	.75
FD1	.15*	.14*	.76
FD2	.56	.20*	.76
FD3	.06*	-.03*	.77
FD4	.25	.27*	.76
FE1	.43	.22*	.76
FE2	.34	.12*	.77
FE3	.49	.21*	.76
FE4	.41	.45	.75
FE5	.33	.40	.75
FF1	.09*	.13*	.76
FF2	.13*	.14*	.76
FF3	.04*	-.04*	.77
FF4	.05*	-.14*	.77
FG1	.24	.35	.75
FG2	.13*	.33	.76
FG3	.44	.37	.75
FG4	.15*	.19*	.76
FG5	.08*	.06*	.77
Durchschnittliche Itemschwierigkeit	.34	Durchschnittliche Itemtrennschärfe	.25

Bemerkung: Die mit * hervorgehobenen Werte stellen Extreme dar, die sowohl den Schwierigkeitsindex als auch die Trennschärfe des Messinstruments beeinträchtigen.

Zur Beurteilung der Trennschärfe des Messinstruments ist zu festzustellen, dass nur knapp die Hälfte der Items den Anforderungen einer mittelmäßigen Trennschärfe zwischen $|.30| \leq r_{it} \leq |.50|$ genügt. Eine hohe Trennschärfe von $r_{it} \geq |.50|$ ist nicht vorzufinden. Weiterhin findet sich bei vereinzelten Items sogar eine negative Trennschärfe. Dies würde bedeuten, dass dieses Item gerade von Schüler*innen, die generell schwach abgeschnitten haben, richtig beantwortet worden ist. Dies könnte vor allem im Raten der Antwort begründet sein. Bei der Beurteilung dieser Werte sei jedoch die Beeinflussung durch den Schwierigkeitsindex zu berücksichtigen. Hohe und niedrige Werte wirken sich nämlich negativ auf die Trennschärfe der Items aus [38]. Dementsprechend lassen sich die negativen Trennschärfen durch einen hohen Schwierigkeitsgrad dieser Items und nicht durch den vorher beschriebenen Zusammenhang erklären. Insgesamt stellt sich die Trennschärfe des Messinstruments als nicht ausreichend dar, wobei besonders die negativen Trennschärfen ins Gewicht fallen. Würden diese drei Items aus der Betrachtung ausgeschlossen werden, um die durchschnittliche Trennschärfe zu erhöhen, so würde sich die Gesamtreliabilität des Messinstruments, nur minimal verbessern. Dies wird aus der Spalte zu Cronbachs-Alpha in Tabelle 6 ersichtlich. Weiterführende Erläuterungen zu Cronbach´s Alpha und den Berechnungen folgen in Kapitel 5.3.2. Aus diesem Grund ist keines der Items aus der Ergebnisbetrachtung ausgeschlossen und alle vorhandenen Daten verwendet worden. Nichtsdestotrotz wird jedoch die nicht ausreichende Trennschärfe in der Beurteilung der Daten berücksichtigt.

5.3. Beurteilung anhand der Hauptgütekriterien

Zur Beurteilung der Glaubwürdigkeit quantitativer Untersuchungen besteht über das Spektrum der Sozial- und Humanwissenschaften bis in die Didaktik der Naturwissenschaften ein breiter Konsens über die drei wesentlichen Hauptgütekriterien Objektivität, Reliabilität und Validität [38, 49, 51, 56, 57]. Bevor diese jedoch nachfolgend einzeln betrachtet werden, muss verdeutlicht werden, wie sich diese Kriterien gegenseitig beeinflussen. Die Objektivität eines Messinstruments stellt die Voraussetzung für Reliabilität und Validität dar, wobei die Validität zusätzlich noch von der Reliabilität beeinflusst wird [57]. Dies wird in Abbildung 3 übersichtlich dargestellt.

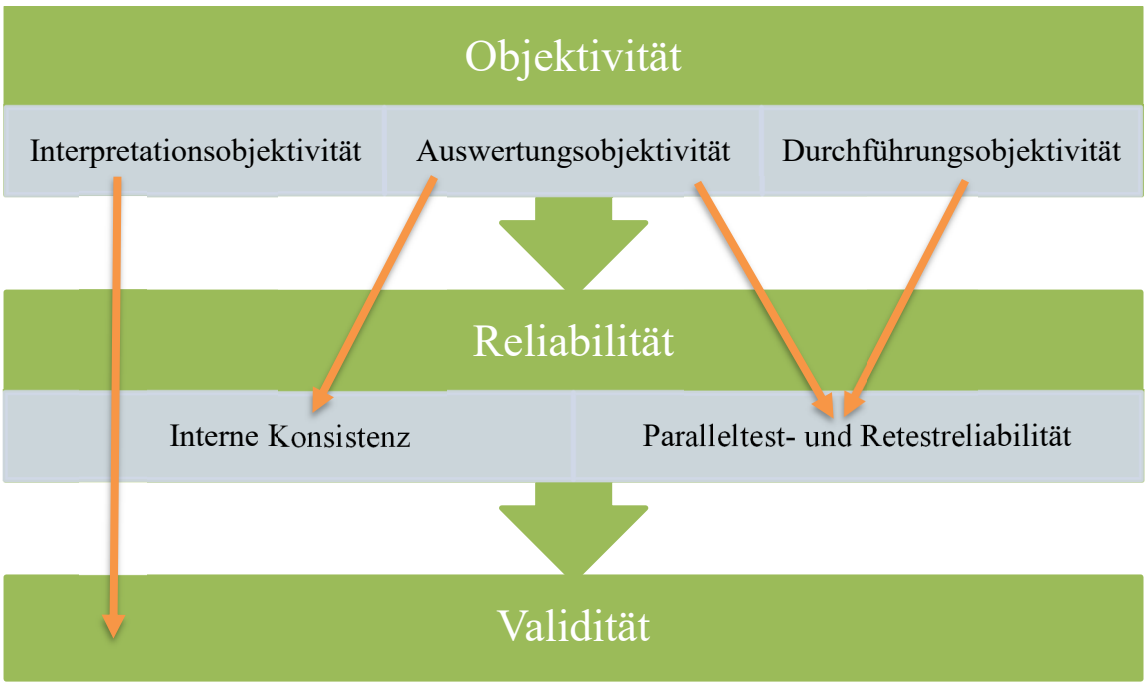


Abbildung 3: Schematische Darstellung der Zusammenhänge zwischen den Hauptgütekriterien [Vgl.48].

5.3.1. Objektivität

Unter der Objektivität eines Messinstruments wird die Reduktion subjektiver Einflüsse auf das Messergebnis verstanden [51]. Das Ziel ist es, möglichst viele Fehlerquellen, die durch Zusatzinformationen über bestimmte Schüler*innen, Sympathie, Voreinstellungen oder erste Eindrücke auf Seiten der Lehrkraft entstehen, zu vermeiden [40]. Aus seiner Beschäftigung mit den *„Stolpersteinen der Diagnostik“* schlussfolgert Höttecke, dass Lehrer*innen all ihre vorangegangenen Interpretationen und Wahrnehmungen für die Leistungsbeurteilung als vorläufig einschätzen und hinterfragen müssen, da diese Beurteilung ansonsten genauso effektiv sei, wie die winterliche Außentemperaturmessung mit einem Fieberthermometer [59]. Da die an der Studie teilnehmenden Schüler*innen dem Testleiter bis zur Durchführung des Pretests unbekannt waren, konnten die beschriebenen Effekte verhindert werden. Jene stellen sich vor allem ein, wenn Lerngruppen über einen gewissen Zeitraum begleitet und näher kennengelernt werden. Nach dem Erhebungszeitraum von sechs Wochen ist nicht davon auszugehen, dass solche Effekte eingetreten sind. Weiterführend ist es möglich, die Objektivität der einzelnen Untersuchungsphasen genauer zu betrachten. Beginnend mit der Durchführungsobjektivität lässt sich feststellen, dass ein standardisiertes Verfahren in Form des Online-Leistungstests mittels LimeSurvey stattgefunden hat und die damit verbundenen Durchführungsbedingungen über alle

Erhebungsgruppen und Messzeitpunkte konstant gehalten werden konnten [57]. Weiterhin waren alle für die Schüler*innen und Lehrkräfte wichtige Instruktionen in schriftlicher Form vor der Erhebung und auch während der Erhebung zu jeder Zeit zugänglich, sodass keine Einflussnahme durch unterschiedliche mündliche Instruktionen auf die Daten stattfand [49]. Anschließend an die bereits festgestellte hohe Durchführungsobjektivität, lässt sich eine ebenfalls hohe Auswertungsobjektivität feststellen. Begründet liegt diese primär in den strengen und gleichzeitig simplen Auswertungsregeln des Single-Choice-Designs des Leistungstest. In Kombination mit der automatischen Auswertung und Aufbereitung der Daten durch LimeSurvey und SPSS konnte der für Fehler und niedrige Objektivität verantwortlichen Faktor Mensch aus der Betrachtung entfernt werden [57]. Abschließend kann über die Interpretationsobjektivität ausgesagt werden, dass die erhobenen Daten der Experimental- und Kontrollgruppen tabellarisch und graphisch dargestellt worden sind, um so den Vergleich zwischen den Gruppen für alle auswertenden Personen möglichst übersichtlich zu gestalten. Dies sorgt für die bestmögliche Steigerung der Interpretationsobjektivität [57]. Somit kann die Objektivität des Messinstruments über alle Teilbereiche hinweg als zufriedenstellend angesehen werden.

5.3.2. Reliabilität

Unter dem Gütekriterium der Reliabilität ist die Verlässlichkeit von Messergebnissen, oder anders ausgedrückt deren Genauigkeit, zu verstehen. Dabei ist nicht von Belang, ob das Instrument überhaupt das misst, was es zu messen vorgibt [57]. Eine hohe Reliabilität zeigt sich vor allem in einer ausreichenden internen Konsistenz des Messinstruments, sowie stabilen Ergebnissen, die aus Messwiederholungen resultieren [51].

Als Maß für die interne Konsistenz eines Messinstruments für intervallskalierte Daten, wie in dieser Untersuchung, wird Cronbachs-Alpha zur Beurteilung der Reliabilität herangezogen. Dieser Wert beschreibt wie stark das Instrument das theoretische Konstrukt einheitlich widerspiegelt. Hierzu wird die durchschnittliche Korrelation zwischen den Items über Formel 2 berechnet, sodass Cronbachs-Alpha einen Wert zwischen 0 und 1 annehmen kann [51].

$$\alpha = \frac{N}{N-1} \cdot \left(1 - \frac{\sum_{i=1}^{N} \hat{\sigma}_i^2}{\hat{\sigma}_x^2}\right)$$

Formel 2: Cronbachs-Alpha [49].

Bemerkung: $\hat{\sigma}_i^2$ = Varianz des Testitems; $\hat{\sigma}_x^2$ = Varianz des Tests, N = Anzahl der Testitems.

Alle hier beschriebenen Berechnungen und resultierenden Werte wurden ebenfalls mit der Statistiksoftware SPSS ermittelt [53, 54]. Über die gesamte Datenmenge erhoben ergibt sich für den entwickelten Leistungstest ein Wert von $\alpha = .76$, welcher für eine solide, aber noch keine gute Reliabilität ($\alpha \geq .80$) spricht [38, 51]. Dieses akzeptable Ergebnis ist auf die Berücksichtigung verschiedener Aspekte in der Testkonstruktion zurückzuführen. Zum einen wird die mit fünf Antwortoptionen relativ hohe Anzahl der Distraktoren und damit verbunden die geringe Wahrscheinlichkeit des Ratens der richtigen Lösung, die Reliabilität gesteigert haben. Zum anderen wirkt sich die mit vier bis fünf ebenfalls hoch angesetzte Zahl der Frageitems pro Itemskala, die ähnliche oder gar identische Kompetenzen zu einem der Themenbereiche prüfen, positiv auf die Reliabilität aus [44]. Weiterhin lässt sich durch das Pre-Posttest-Design der Vergleich von Cronbachs-Alpha Werten zu verschiedenen Messzeitpunkten in identischen Versuchsgruppen heranziehen. Dabei ist davon auszugehen, dass das zu messende Konstrukt von den Schüler*innen im Pretest noch nicht so gut verstanden wird wie im Posttest, da weitere Unterrichtsstunden zwischen den Tests die Kompetenzen der Schüler*innen steigern sollten. Dies müsste dann dazu führen, dass im Nachtest, auch wenn die Schüler*innen nicht raten sollten, dennoch im Schnitt weniger geraten wird als noch im Pretest. Als Folge daraus sollte sich die innere Konsistenz und damit Cronbachs-Alpha erhöhen [51]. Dieses Phänomen ist über alle Versuchsgruppen mit Ausnahme der Experimentalgruppe an der St. Ursula-Schule zu beobachten. Die dazugehörigen Daten finden sich in Tabelle 7.

Tabelle 7: Cronbachs-Alpha Werte als Maß für die Reliabilität des Leistungstest in allen Versuchsgruppen.

Versuchsgruppe	UVK	UNK	UVE	UNE	BVK	BNK	BVE	BNE
Cronbachs-Alpha	.64	.74	.82	.76	.64	.79	.66	.87

Bemerkung: Zur Abkürzung der Versuchsgruppen wurde folgende Codierung genutzt: **B** Bismarckschule; **U** St. Ursula-Schule; **V** Pretest; **N** Posttest; **E** Experimentalgruppe; **K** Kontrollgruppe.

Ein weiteres Maß für die Reliabilität stellt die Retestmethode dar, bei der zu zwei aufeinanderfolgenden Messzeitpunkten die dasselbe Konstrukt in derselben Gruppe gemessen wird [57]. Dieses Vorgehen erinnert stark an das dieser Untersuchung zugrundeliegende Pre-Posttest-Design. Anhand der Korrelation der Messwerte aus Pre- und Posttest lässt sich eine Aussage über die Stabilität der Messergebnisse treffen und damit die Reliabilität

des Messinstruments beurteilen [57]. Aus der Tabelle 8 wird ersichtlich, dass hohe positive Korrelationen mit einem hervorragenden Signifikanzniveau für die Experimental und Kontrollgruppe an der St. Ursula-Schule vorzufinden sind.

Tabelle 8: Korrelation nach Pearson für die vergleichbaren Leistungen aller Versuchsgruppen in den Pre- und Posttest.

		UNE	UNK	BNE	BNK
UVE	Pearson-Korrelation	.622*	-.776**	-.474	.205
	Signifikanz (2-seitig)	.013	.005	.075	.482
	N	15	11	15	14
UVK	Pearson-Korrelation	-.621*	.717*	.531	-.164
	Signifikanz (2-seitig)	.041	.013	.093	.630
	N	11	11	11	11
BVE	Pearson-Korrelation	-.002	.260	.122	.417
	Signifikanz (2-seitig)	.996	.440	.642	.138
	N	15	11	17	14
BVK	Pearson-Korrelation	-.461	-.423	.055	-.462
	Signifikanz (2-seitig)	.097	.194	.853	.096
	N	14	11	14	14

Bemerkung: * Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant. ** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Dies lässt auf eine große Stabilität der Messergebnisse schließen und damit lässt sich die mittels Cronbach-Alpha ermittelte solide Reliabilität bestätigen. In der Experimentalgruppe der Bismarckschule hingegen findet sich nur noch eine geringe positive Korrelation zwischen den Pre- und Posttestleistungen mit einem nicht ausreichenden Signifikanzniveau. In der Kontrollgruppe der Bismarckschule kann eine mittlere negative Korrelation vorgefunden werden, die vermutlich auf klassenarbeitsbedingte Lerneffekte der Schüler*innen zwischen Pre- und Posttest zurückzuführen ist. Genaueres dazu folgt in Kapitel 7. Resümierend lässt sich also festhalten, dass das entwickelte Messinstrument, trotz kleinerer Schwächen eine ausreichende Reliabilität vorweist.

5.3.3. Validität

Die Validität eines Tests trifft eine Aussage darüber, inwieweit dieser Test tatsächlich das misst was er messen soll oder zu messen vorgibt [57]. Dieses Gütekriterium gibt folglich an, ob das Messinstrument überhaupt geeignet ist, das theoretische Konstrukt in angemessener Weise abzubilden, ähnlich zu einem Stromstärkemessgerät mit welchem sich

eben keine Spannungen messen lassen [51]. Wie bei der Objektivität wird auch die Validität in drei wesentliche Bereiche unterteilt: Inhalts-, Konstrukt- und Kriteriumsvalidität. Beginnend mit der Inhaltsvalidität wird festgelegt, ob das Instrument das Konstrukt inhaltlich ausreichend abbildet. Dies lässt sich jedoch schlecht in einen numerischen Wert fassen, weshalb meist Expertenurteile herangezogen werden müssen [57]. Für die Entwicklung dieses Messinstruments wurde sich, wie bereits mehrfach erwähnt, nach den inhalts- und prozessbezogenen Kompetenzen des niedersächsischen Kerncurriculums gerichtet. Ferner sind Expertenurteile in Form eines Feedbacks der unterrichtenden Fachlehrkräfte eingeholt und deren Anmerkungen in den Leistungstest eingearbeitet worden. Hiermit sollte eine ausreichende Inhaltsvalidität vorherrschen.

Die Konstruktvalidität soll widerspiegeln, ob das Messinstrument tatsächlich das gemessen hat, was es messen sollte. Hierzu wurde eine explorative Faktorenanalyse in SPSS durchgeführt. In diesem Verfahren werden Faktoren, oder Kategorien ermittelt, denen die einzelnen Items inhaltlich anhand der empirischen Daten zugeordnet werden können [57]. Stimmen diese Faktoren mit den bereits beschriebenen sieben Themenbereichen des Leistungstests überein, so kann von einer hohen Konstruktvalidität ausgegangen werden. Tabelle 9 stellt die Ergebnisse der explorativen Faktorenanalyse dar. Es wird ersichtlich, dass hohe Korrelationen ($r \geq |.50|$) nur in den wenigsten Fällen vorhanden sind. Damit eine zufriedenstellende Konstruktvalidität des Messinstruments hätte erreicht werden können, müssten einzelne Fragegruppen, z.B. FA1-FA4, auf nur einen Faktor hoch korrelieren und alle anderen Items sollten möglichst niedrig bis gar nicht auf diesen Faktor korrelieren. Für die Items FA1-FA4 lässt sich noch eindeutig der Faktor 4 als Bestätigung der in den theoretischen Betrachtungen entwickelten Themenbereiche des Leistungstests heranziehen. Es wird jedoch recht zügig ersichtlich, dass diese klare Differenzierung anhand der Daten aus Tabelle 9 für alle anderen Items nicht möglich ist und einige von diesen sogar auf keinen der sieben Faktoren mit ausreichender Stärke laden (FB2, FB4, FC1, FD1, FG4, FG5). Auch wenn mittels eines Kaiser-Meyer-Olkin- und zusätzlichem Bartlett-Test bestätigt werden konnte, dass die Daten für eine Faktorenanalyse geeignet sind, lieferte die Faktorenanalyse selbst aufgrund nicht ausreichender Signifikanz ($p = .43$) keine aussagekräftigen Ergebnisse. Verschiedenste Anpassungen, wie die Reduktion der Faktoren oder der Ausschluss einzelner Items lieferten keine nennenswerten Verbesserungen. Lediglich die Reduktion auf nur noch einen verbleibenden Faktor steigerte die Aussagekraft der Analyse auf ein signifikantes Niveau. Daher ist davon auszugehen, dass die Konstruktvalidität des Leistungstests als nicht ausreichend angesehen werden muss.

Eine weiterführende Betrachtung dieser Problematik erfolgt in den Kapiteln 7 und 8 dieser Arbeit.

Tabelle 9: Rotierte Korrelationsmatrix der explorativen Faktorenanalyse.

Faktor	1	2	3	4	5	6	7
FA1			.51	.31			
FA2				.40			
FA3				.79			
FA4	.40			.44			
FB1	.40						
FB2							
FB3			.43				
FB4							
FC1							
FC2	.56						
FC3	.35						
FC4		.35	.54				
FD1							
FD2		.39					
FD3							
FD4	.44						
FE1	.41						
FE2		.36					
FE3		.47					
FE4		.58					
FE5						.79	
FF1					.57		
FF2					.59		
FF3							.78
FF4		-.44					
FG1	.38						
FG2		.33			.43		
FG3			.31			.50	
FG4							
FG5							

Bemerkung: Es werden nur aussagekräftige Faktorladungen ($r \geq |.30|$) dargestellt.

Als letzten Teil der Analyse wird die Kriteriumsvalidität, also die Stärke des Zusammenhangs zwischen den Ergebnissen des Leistungstest und einem wichtigen externen Merkmal, in diesem Fall der Physiknote aus dem vorangegangenen Schulhalbjahr, ermittelt [57]. Konkret wurde der Korrelationskoeffizient nach Pearson (s. Formel 1) zwischen den im Pretest erhobenen Noten und der jeweils erreichten Punktzahl auf einen mittelstarken Wert von $r = .48$ ($p < .01$) bestimmt. Dieses Ergebnis lässt auf einen substantiellen Zusammenhang zwischen der Leistung im Pretest und den bisherigen Physiknoten schließen und bescheinigt eine solide Kriteriumsvalidität des Messinstruments [46].
Abschließend lässt sich festhalten, dass das entwickelte Erhebungsinstrument, abgesehen von der Konstruktvalidität, objektive, reliable und valide Ergebnisse liefern kann.

5.4. Entwicklung des Feedback-Fragebogens zum EXIT-Spiel

Neben dem eigentlichen Forschungsinteresse dieser Arbeit, wurde in den Experimentalgruppen beider Schulen nach Beendigung des Spiels eine zusätzliche Erhebung über die Wahrnehmung des Spieleinsatzes durch die Schüler*innen durchgeführt. Der dazu verwendete Feedbackfragebogen soll hier kurz erläutert werden. Im Gegensatz zum Leistungstest handelte es sich bei diesem zweiten Erhebungsinstrument um einen klassischen analogen Fragebogen. Das komplette Design und auch die zur Erhebung verwendeten Items sind dabei an die von Wilde aufgestellte Kurzskala zur Messung intrinsischer Motivation angelehnt [33]. Da dieser die Güte des Messinstruments bereits ausführlich betrachtet hat, wurde von einer erneuten Analyse abgesehen [33].
Anhand der in zum Abschluss von Kapitel 2.4 beschriebenen Feststellungen von Grassinger [32] hat Wilde ein Messinstrument entwickelt, mit welchem sich die intrinsische Motivation von Schüler*innen zuverlässig messen lässt [33]. Da dieses Messinstrument für einen Museumsbesuch und nicht für den Einsatz von Spielen im Unterricht konzipiert worden ist, wurde aufgrund der vielen Gemeinsamkeiten ein adaptierter Fragebogen zur Motivationsmessung für diese Untersuchung entwickelt (s. Anhang). Zu besagten Gemeinsamkeiten zählt unter anderem die Autonomie und Freiheit der individuellen Gestaltung des Fortgangs des Spiels durch die Schüler*innen selbst. Da sie frei von Zwängen des Leistungsdrucks und der Benotung sind, können sich voll und ganz auf ihren individuellen Lernprozess konzentrieren. Diese Freiheit bedingt automatisch eine hohe Anforderung an die Selbstregulation und Organisation der Schüler*innen. Der Lernzugang dagegen verschiebt sich durch die spielerische Handlung von einem kognitiv-orientierten zu einem emotional-affektiven Sinn [33]. Die daraus resultierenden zwölf Items decken

die vier wesentlichen Bereiche Interesse und Vergnügen, Kompetenz, Wahlfreiheit und Druck, der intrinsischen Motivation ab, sind jedoch anders als in Wildes Version zum Teil in reduzierter Anzahl vorhanden oder durch andere auf das Spiel angepasste Items, z. B. die Wahrnehmung der Gruppengröße, ersetzt worden (s. Anhang). Neben der Messung der intrinsischen Motivation der Schüler*innen dient das Instrument auch als Werkzeug zur Erfassung von potentiell problematischen Eigenschaften des Spiels, damit diese im Anschluss an die Untersuchung ebenfalls beseitig werden können und damit das Spiel für zukünftige Einsätze im Unterricht verbessert werden kann. Der vollständige Fragebogen ist dem digitalen Anhang zu entnehmen. In der nachfolgenden Tabelle 10 sind die Frageitems dargestellt.

Tabelle 10: Fragebogen zur Messung der intrinsischen Motivation der Schüler*innen durch den Spieleinsatz [33].

Item-Code	**Frageitem**	++	+	**o**	**-**	--
FM1	Die Durchführung des Spiels hat mir Spaß gemacht.					
FM2	Ich fand die Inhalte des Spiels interessant.					
FM3	Während des Spiels fühlte ich mich unter Druck gesetzt.					
FM4	Die Rätsel und Fragen des Spiels haben mich überfordert.					
FM5	Ich konnte mein bisheriges Wissen gut im Spiel anwenden.					
FM6	Ich habe durch das Spiel etwas Neues gelernt.					
FM7	Das Spiel hat mir meine Wissenslücken aufgezeigt.					
FM8	Das Spiel hat mein Interesse für die Kernphysik verstärkt.					
FM9	Ich habe während des gesamten Spiels aktiv in meiner Gruppe mitgearbeitet und zur Lösung der Rätsel beigetragen.					
FM10	Ich fand die Gruppengröße während des Spiels angemessen.					
FM11	Das Spiel hat mich stärker motiviert als konventioneller Physikunterricht.					
FM12	Ich möchte, dass zukünftig öfter Spiele im Unterricht als Lernmethode eingesetzt werden.					

6. Ergebnisse

6.1. Durchführung des EXIT-Spiels „ESCAPE – Gefangen bei der Wismut“

Bevor die Ergebnisse der Leistungstests dargestellt werden, soll in diesem Kapitel der Ablauf der Spielstunden, sowie Beobachtungen der Versuchsleitung aus den Experimentalgruppen beider Schulen genauer beschrieben werden. Alle nachfolgend erwähnten Rätselmaterialien sind dem digitalen Anhang dieser Arbeit zu entnehmen.

Der Faktor Bearbeitungszeit stellte sich im Laufe der Spielstunden als kritischste Komponente der Erhebung dar. Die Anpassungen des Spiels (s. Kapitel 5.1) sollte die notwendige Reduktion der Spielzeit auf lediglich zwei Unterrichtsdoppelstunden sicherstellen. Jeweils die leistungsstärkere Gruppe pro Klasse hat es in der vorgegebenen Zeit geschafft, das Spiel erfolgreich zu beenden. Die leistungsschwächeren Gruppen haben bis zum Ablauf der Spielzeit weitergearbeitet und konnte beide das vorletzte Rätsel des Spiels nicht mehr vollständig beenden. Zum erfolgreichen Abschluss des Spiels hätten diese beiden Gruppen noch circa 15 Minuten benötigt. Da bereits 30 Minuten der ersten Spielstunde für das Lesen der Spielanleitung verwendet wurden, könnte eine Zeiteinsparung an dieser Stelle dafür sorgen, dass alle Gruppen des Leistungsspektrums das Spiel in der vorgegebenen Bearbeitungszeit erfolgreich beenden können. Ein alternatives Beschäftigungsprogramm für Gruppen, die das Spiel vorzeitig beenden, sollte ebenfalls entwickelt werden, um die verbleibende Lernzeit so effektiv wie möglich zu nutzen. Denkbar wäre hier eine Nachbesprechung des Spiels zwischen der unterrichtenden Lehrkraft und den Schüler*innen der Spielgruppe. Sanchez Untersuchungsergebnisse belegen einen wesentlichen Einfluss von Nachbesprechungen auf den Kompetenzzuwachs der Schüler*innen und stellen deren Bedeutung für den erfolgreichen Einsatz von EXIT-Spielen im Unterricht heraus [37].

Die Zeitproblematik stellte sich vornehmlich durch den nicht den Erwartungen entsprechenden Einsatz von Hilfe-Karten dar. Bereits während der Entwicklung des Spiels wurde davon ausgegangen, dass die Schüler*innen, sofern diese ein Rätsel nicht lösen konnten, mittels Hilfe-Karten eigenständig Unterstützung heranziehen. Hierbei wurde jedoch der Wettkampfcharakter innerhalb der Klassen unterschätzt. Jede Spielgruppe versuchte schneller als die andere Gruppe und mit einem möglichst besseren Ergebnis abzuschließen. Da die Benutzung der Hilfe-Karten einen Punktabzug impliziert, wurden diese nahezu überhaupt nicht verwendet. Als Folge daraus ging ein großer Teil der Arbeitszeit aufgrund des internen Wettkampfes verloren. Für den zukünftigen Einsatz des Spiels

sollte diese negative Wahrnehmung der Hilfe-Karten berücksichtigt und eine Abwandlung des Punktesystems in Betracht gezogen werden.
Während der Durchführung stellten sich einzelne Rätsel als besonders zeitintensiv und fordernd heraus. Hierzu zählte das fünfte Rätsel zur Karlsruher Nuklidkarte. Besonders die dort verwendeten römischen Zahlen und griechischen Buchstaben sorgten dafür, dass die Schüler*innen das Rätsel als zeitintensiv und wenig spannend wahrgenommen haben. Das vierte Rätsel zur γ-Strahlung hingegen benötigte über alle Gruppen hinweg einen ungefähr gleich hohen Zeitbedarf, wurde jedoch von den Schüler*innen aufgrund des zu entschlüsselnden Musters in einem Brief als positiv beschrieben. Das wohl herausforderndste Rätsel wurde von allen Schüler*innen in den Aufgaben zu den natürlichen Zerfallsreihen (Rätsel 8) gesehen. Trotz expliziter Übungsphasen zum 4n-Schema im vorangegangenen Unterricht benötigten an dieser Stelle alle Gruppen Unterstützung in Form von Hilfe-Karten, um das Rätsel zu lösen und im Spiel fortzufahren. Auch hier zeigte sich die ablehnende Haltung gegenüber der Hilfe-Karten, da die Schüler*innen vornehmlich versuchten, Hinweise und Hilfestellungen von der Versuchsleitung in Erfahrung zu bringen, um den Punktabzug durch die Hilfe-Karten zu vermeiden.
Besonders zu Beginn der ersten Spielstunde häuften sich über alle Gruppen hinweg Nachfragen zum Codierungssystem. Den Schüler*innen wurde anhand der Spielanleitung und zusätzlichen Erläuterungen durch die Versuchsleitung nicht ersichtlich, wie sie nach einem gelösten Rätsel die passende Lösungskarte finden konnten. Besonders die Rolle der zu den Logikrätseln gehörenden Multiple-Choice-Fragen, welche sich auf zusätzlichen Spielkarten befanden, stellte die Schüler*innen zu Beginn vor große Probleme. Ihnen wurde nicht ersichtlich, dass die Kombination der Ergebnisse aus dem Logikrätsel und der dazugehörenden Multiple-Choice-Frage den Namen der Lösungskarte lieferte. Aus diesem Grund sollte über eine Überarbeitung der Spielanleitung oder eine generelle Vereinfachung des Codierungssystems nachgedacht werden.
Abschließend lässt sich bezüglich der benötigten Spielzeit festhalten, dass die gewünschte Verkürzung zumindest für die leistungsstärkeren Schüler*innen erfolgreich verlaufen ist. Die beschriebenen Beobachtungen zur Wahrnehmung einzelner Rätsel können für zukünftige Einsätze zur Verbesserung des Spiels herangezogen werden. Weiterführende Informationen zur Wahrnehmung des Spiels durch die Schüler*innen, welche mittels des entwickelten Fragebogens erhoben wurden (s. Kapitel 5.4), folgen in den Kapiteln 6.2.3 sowie 6.3.3 dieser Arbeit.

6.2. Bismarckschule

6.2.1. Experimentalgruppe

Im Pretest haben die Schüler*innen der Experimentalgruppe ein durchschnittliches Ergebnis von 10,4 Punkten bei einer Standardabweichung von 3,8 Punkten erreicht, wohingegen sich der Mittelwert im Posttest geringfügig auf 10,1 Punkte reduzierte und sich die Standardabweichung auf 6,0 Punkten steigerte. Die erreichten Punktzahlen erscheinen bei maximal 30 möglichen Punkten recht niedrig, spiegeln jedoch auf Grund vergleichbarer Mittelwerte das tatsächliche Leistungsniveau aller vier Versuchsklassen wieder, wobei die in Kapitel 5.2.4 beschriebe Schwierigkeit der Items natürlich berücksichtigt werden muss.

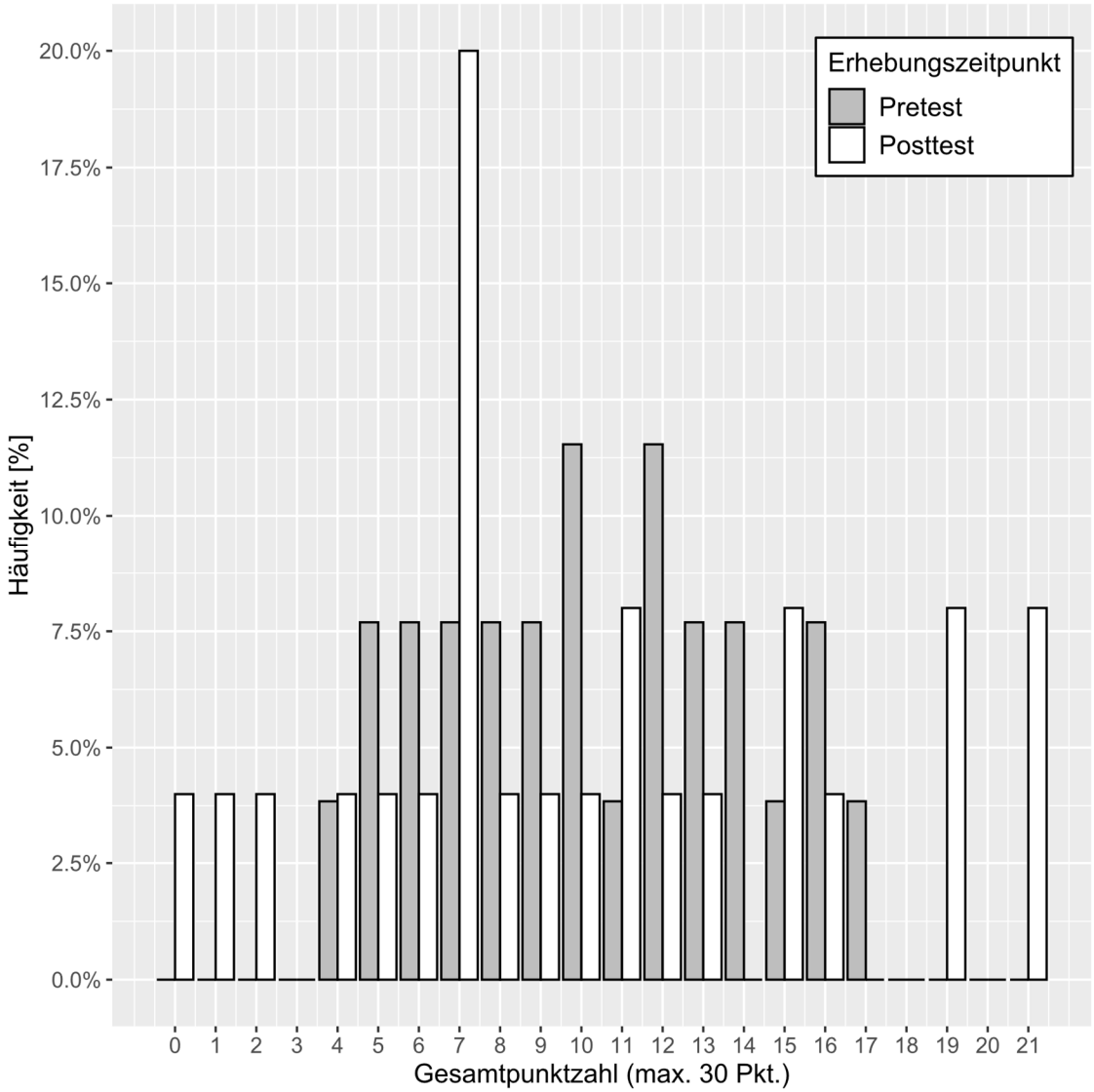


Abbildung 4: Punkteverteilung im Pre- & Posttest der Experimentalgruppe an der Bismarckschule (BE).

Abbildung 4 stellt die prozentuale Häufigkeitsverteilung der Gesamtpunktzahlen zu beiden Messzeitpunkten dar. Eine zu erwartende Normalverteilung der Messwerte um den

Mittelwert lässt sich im Diagramm nur bedingt erkennen. Weiterhin wird ersichtlich, dass die Messwerte im Posttest, verglichen zum Pretest, stärker zu Extremwerten tendieren. Die Streuung der Ergebnisse hat sich vom ersten zum zweiten Messzeitpunkt von 13 auf 21 von 30 möglichen Punkten erhöht. Es wurden sowohl mehr extrem geringe als auch extrem hohe Punktzahlen, bezogen auf die anderen Schüler*innen, gemessen. Der in der Hypothese 1 formulierte positive Effekt des Spiels auf die Ergebnisse des Posttests lässt sich hier nicht anhand einer deutlichen Verschiebung der Häufigkeitsverteilung hin zu höheren oder zumindest deutlich zweistelligen Punktzahlen und einem signifikant erhöhten Mittelwert beobachten. Ebenfalls ergab ein mittels SPSS durchgeführter t-Test für gepaarte Stichproben, dass der Unterschied der Mittelwerte beider Erhebungszeitpunkte statistisch nicht signifikant war (p = .79) [54]. Der mittels SPSS bestimmte p-Wert liegt oberhalb des festgelegten Signifikanzniveaus von 0.05, weshalb die Nullhypothese beibehalten wird. Für den t-Test in SPSS wurde folgende Formel verwendet:

$$t = \frac{\overline{D}}{\frac{S_D}{\sqrt{N}}}$$

Formel 3: t-Test für gepaarte Stichproben [54].

Bemerkung: $\overline{D}$ = Mittelwert der Punktedifferenz zwischen beiden Messzeitpunkten; S_D = Standardabweichung der Stichprobe

Zusätzlich wurde zur Quantifizierung eines möglichen Effekts des Spieleinsatzes Cohen´s d als Maß für die Effektstärke des Spieleinsatzes herangezogen. Dieses ergibt sich aus der Mittelwertdifferenz beider Messzeitpunkte, welche anschließend durch die Standardabweichung dividiert wird [53]. Es wird von kleinen ($d \geq .20$), mittleren ($d \geq .50$) und großen ($d \geq .80$) Effekten gesprochen [59]. In SPSS wurde Cohen´s d über folgende Formel bestimmt:

$$d = \frac{\bar{x} - \bar{y}}{S_D}$$

Formel 4: Effektstärkemaß Cohen´s d [38].

Mit einem berechneten Wert von d = .06 wurde damit kein merklicher Effekt des Spieleinsatzes auf die Leistung im Posttest nachgewiesen.

Neben generellen Aussagen zur Gesamtheit der Kontrollgruppe kann auch die individuelle Leistung einzelner Schüler*innen auf Grund standardisierter Teilnehmenden-Codes betrachtet werden. Die Stichprobengröße (N = 18) reduzierte sich hierfür auf Schüler*innen, deren Pre- und Posttestleistungen anhand ihres Teilnehmenden-Codes miteinander

verknüpft werden konnten. Diese individuellen Leistungsunterschiede sind in Abbildung 5 dargestellt.

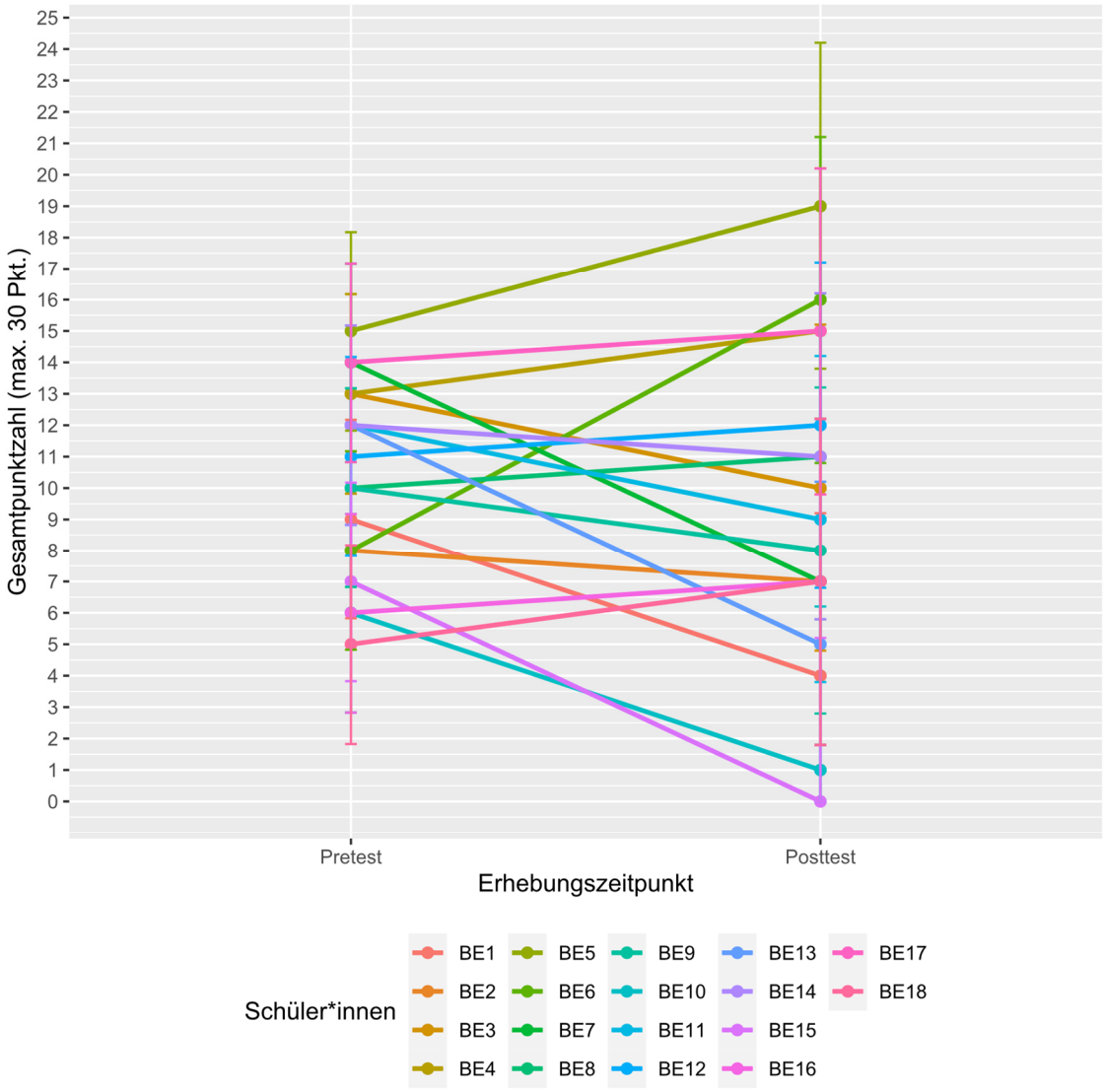


Abbildung 5: Vergleich der Individualleistung am Pre- und Posttest teilnehmender Schüler*innen aus der Versuchsgruppe BE.

Die dargestellten Fehlerbalken veranschaulichen die Standardabweichung zum jeweiligen Messzeitpunkt (Pretest: S_D = 3,2 Pkt.; Posttest: S_D = 5,2 Pkt.). Aus Gründen des Datenschutzes sind die originalen Teilnehmenden-Codes in der Legende durch, der bisherigen Kodierung entsprechende, Kürzel ersetzt worden. Bei der Betrachtung von Abbildung 5 fällt auf, dass etwas mehr als die Hälfte dieser Schüler*innen sich im Posttest, verglichen zu Pretest verschlechtert haben. Neben dem Gesamtergebnis über den vollständig bearbeiteten Leistungstest erlaubten die Daten auch einen Vergleich von einzelnen Frageitems zwischen Pre- und Posttest. Dieser ergab ebenfalls mittels t-Test Über-

prüfung keine signifikanten Änderungen zwischen der Häufigkeit richtig oder falsch beantworteter einzelner Frageitems, die auf einen Effekt des Spiels schließen lassen könnten.

6.2.2. Kontrollgruppe

Die durchschnittlichen Punktzahlen in der Kontrollgruppe der Bismarckschule belaufen sich zum ersten Erhebungszeitpunkt auf 10,6 Punkte mit einer Standardabweichung von 3,7 Punkten und erhöhten sich zum Posttest auf 13,1 Punkte mit einer Standardabweichung von 4,8 Punkten.

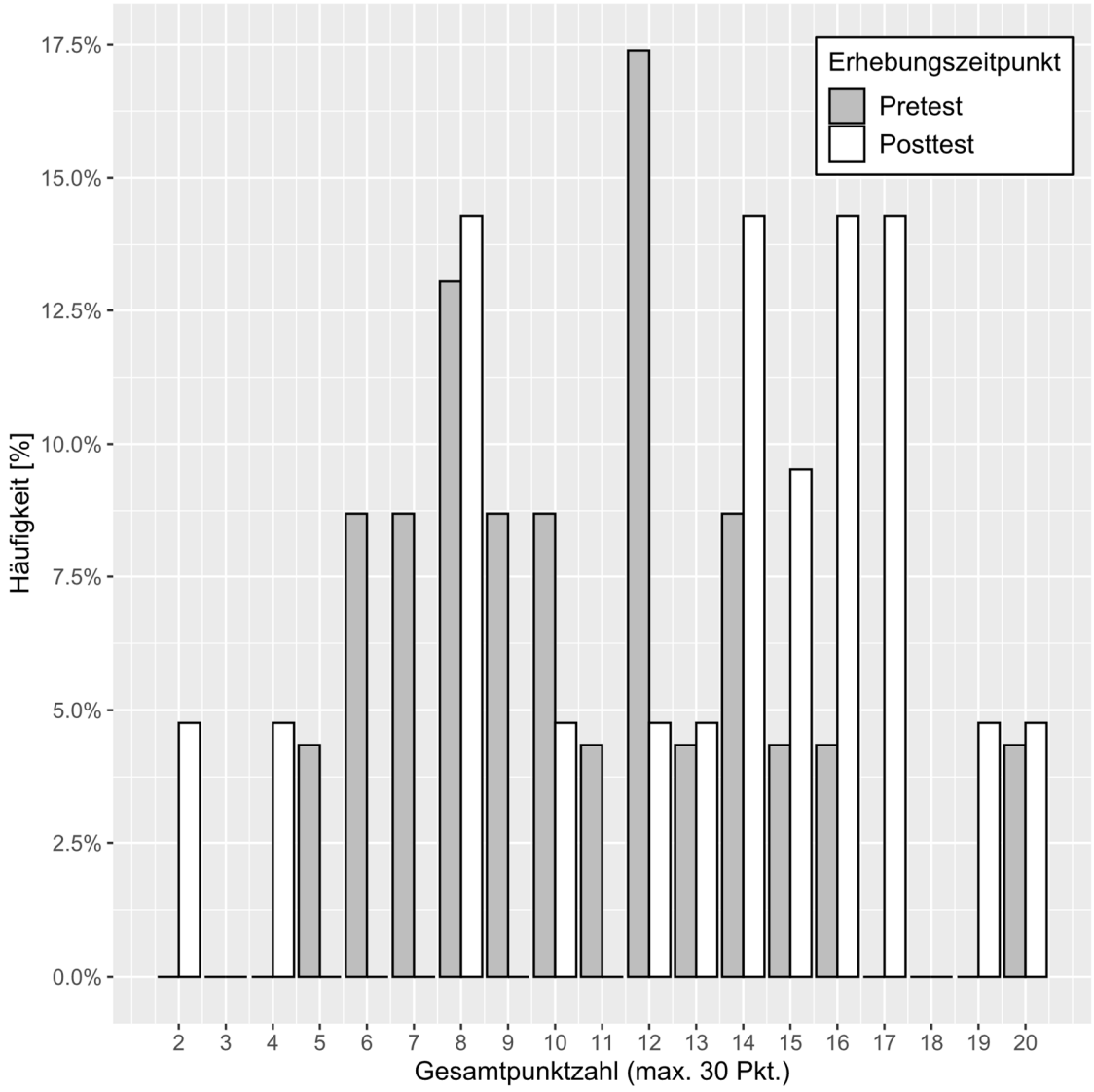


Abbildung 6: Punkteverteilung im Pre- & Posttest der Kontrollgruppe an der Bismarckschule (BK).

Die Abbildung 6 stellt hier wieder die Häufigkeitsverteilung der Gesamtpunktzahlen dar, aus welcher in diesem Fall eine merkliche Verschiebung der Ergebnisse zwischen Pre-

und Posttest von einstelligen Werten hin zu mehrheitlich zweistelligen Punktzahlen ersichtlich wird. Weiterhin wurde für die Beurteilung der Signifikanz des Unterschieds auch hier wieder ein t-Test durchgeführt (p = .13), welcher ebenfalls die Verwerfung der Nullhypothese nicht ermöglicht. Eine Abschätzung der Effektstärke (d = -.34) lässt hier einen mittelstarken negativen Zusammenhang zwischen den Leistungen im Pre- und Posttest vermuten. Die Streuung der Messwerte steigt auch hier leicht von 15 auf 18 Punkte an.

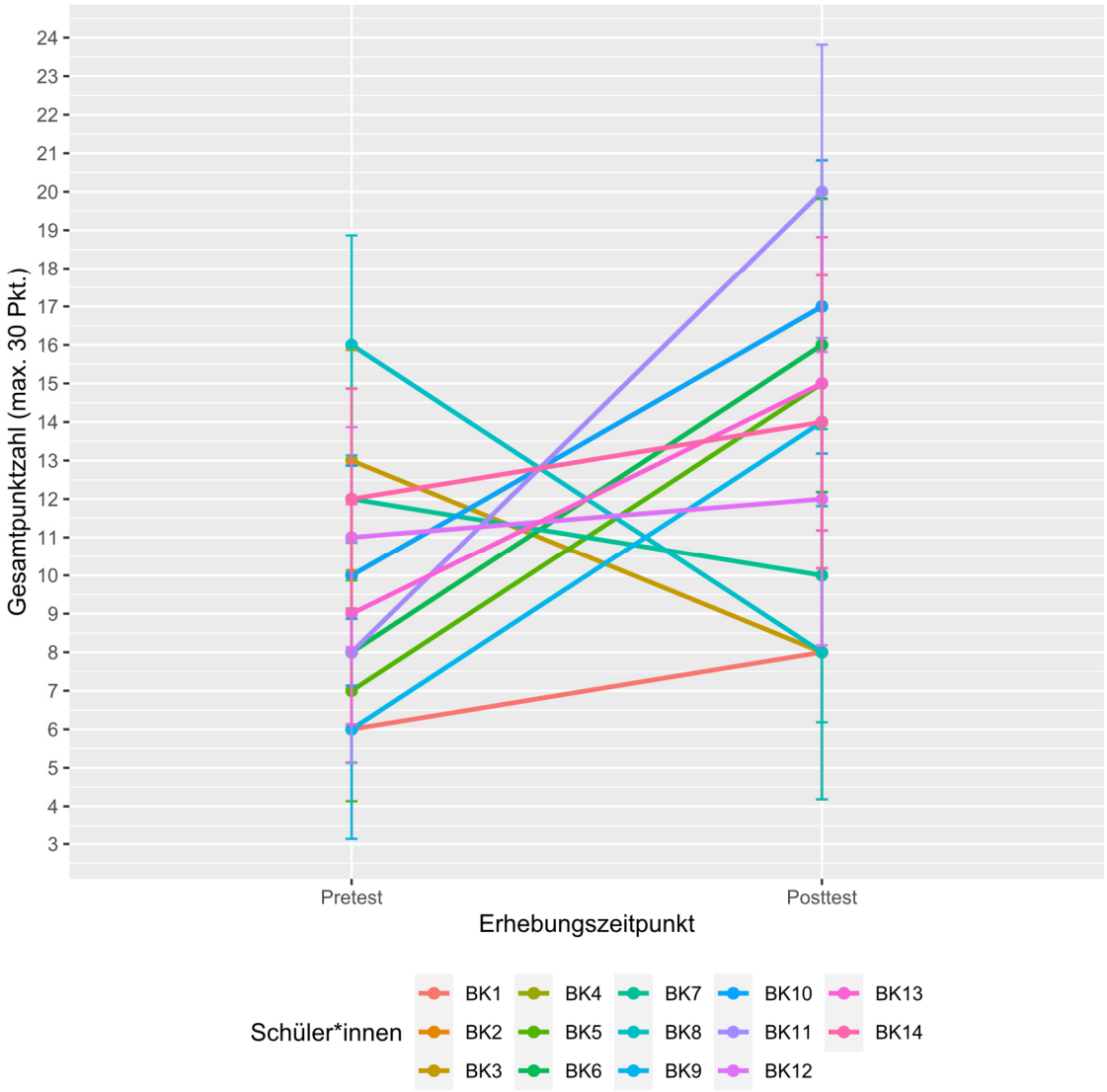


Abbildung 7: Vergleich der Individualleistung am Pre- und Posttest teilnehmender Schüler*innen aus der Versuchsgruppe BK.

Für die individuelle Betrachtung konnten anhand der Teilnehmenden-Codes in dieser Lerngruppe ebenfalls weniger Schüler*innen herangezogen werden (N = 14), deren Leistungsentwicklung in Abbildung 7 dargestellt werden (Pretest: S_D = 2,9 Pkt.; Posttest: S_D = 3,8 Pkt.). Aus dieser Grafik wird ersichtlich, dass die große Mehrheit der Schüler*innen

ihre Individualleistung vom Pre- zum Posttest verbessern konnten, wobei einige davon, z. B. BK11, die Punktzahl auch über das Fehlerniveau hinaus steigern konnten. Trotz der unter den Schüler*innen breit gefächerten Leistungssteigerung sind auch einige Verschlechterungen der Gesamtpunktzahlen erkennbar, die im Vergleich zur Experimentalgruppe jedoch weniger stark vertreten sind.

Auch hier wurden die Häufigkeiten der richtig beantworteten Frageitems zwischen beiden Messzeitpunkten verglichen. Der t-Test lieferte nur für das Item FA4 ($p \leq .05$) signifikante Steigerung der Mittelwerte zwischen Pre- und Posttest.

6.2.3. Wahrnehmung des Spiels durch die Schüler*innen

Über die durchgeführten Leistungstest hinaus wurden auch Daten zur Wahrnehmung des Spieleinsatzes in den Experimentalgruppen erhoben. Die Schüler*innen hatten dazu im Anschluss an die Spieldurchführung die Gelegenheit, anhand eines Fragebogens bestehend aus 12 Items (s. Kapitel 5.4) mittels einer fünfgliedrigen Skala ihre Wahrnehmung von sehr positiv (++) bis hin zu sehr negativ (- -) auszudrücken. Für die jetzt stattfindenden Auswertung mussten diese Daten Likert-skaliert werden. Hierzu wurden den symbolischen Zuordnungen von ++ bis - - die numerischen Werte 5 (++) bis 1 (- -) zugeordnet [38]. Anhand dieser Skalierung war die Erstellung von Abbildung 8 möglich, in welcher die Zustimmungswerte zu jedem der 12 Frageitems inklusive deren Standardabweichung als Fehlerbalken aufgetragen sind.

Bei der Betrachtung dieser Ergebnisse ist vor allem der Wunsch der Schüler*innen nach einem häufigeren Einsatz von Spielen im Physikunterricht (FM12) zu erkennen. Diese Beobachtung geht einher mit der ebenfalls relativ hohen Motivation, die das Spiel bei den Schüler*innen ausgelöst hat (FM1 & FM11). Damit spiegeln sich die in den theoretischen Vorbetrachtungen erwähnten empirischen Ergebnisse bezüglich des Spiels als motivationssteigernde Maßnahme auch in diesen Daten wieder [11, 19, 20, 24, 34]. Ebenfalls hat das Spiel die Schüler*innen in nur seltenen Fällen unter Druck gesetzt (FM3), was als weiterer Indikator für eine durch das Spiel hervorgerufene intrinsische Motivation herangezogen werden kann [33]. Jedoch sei anzumerken, dass die Schüler*innen zwar zur Durchführung des Spiels motiviert gewesen sind, gleichzeitig haben diese jedoch auch angegeben, dass das Spiel ihr Interesse für die Kernphysik nicht zusätzlich gesteigert hat (FM2 & FM8). So scheint sich der motivatorische Charakter des Spiels nicht positiv auf das generelle Ansehen und Interesse für die Physik ausgewirkt zu haben. Die aus den

Daten der Leistungstests hervorstechende Tatsache, dass es zu keinem signifikanten Lernzugewinn durch die Anwendung des Spiels gekommen ist, wurde nur bedingt von den Schüler*innen so wahrgenommen (FM 6). Diese Abweichungen zwischen Datenlage und Schüler*innenwahrnehmung wird auch anhand der Spielgruppengröße deutlich (FM10).

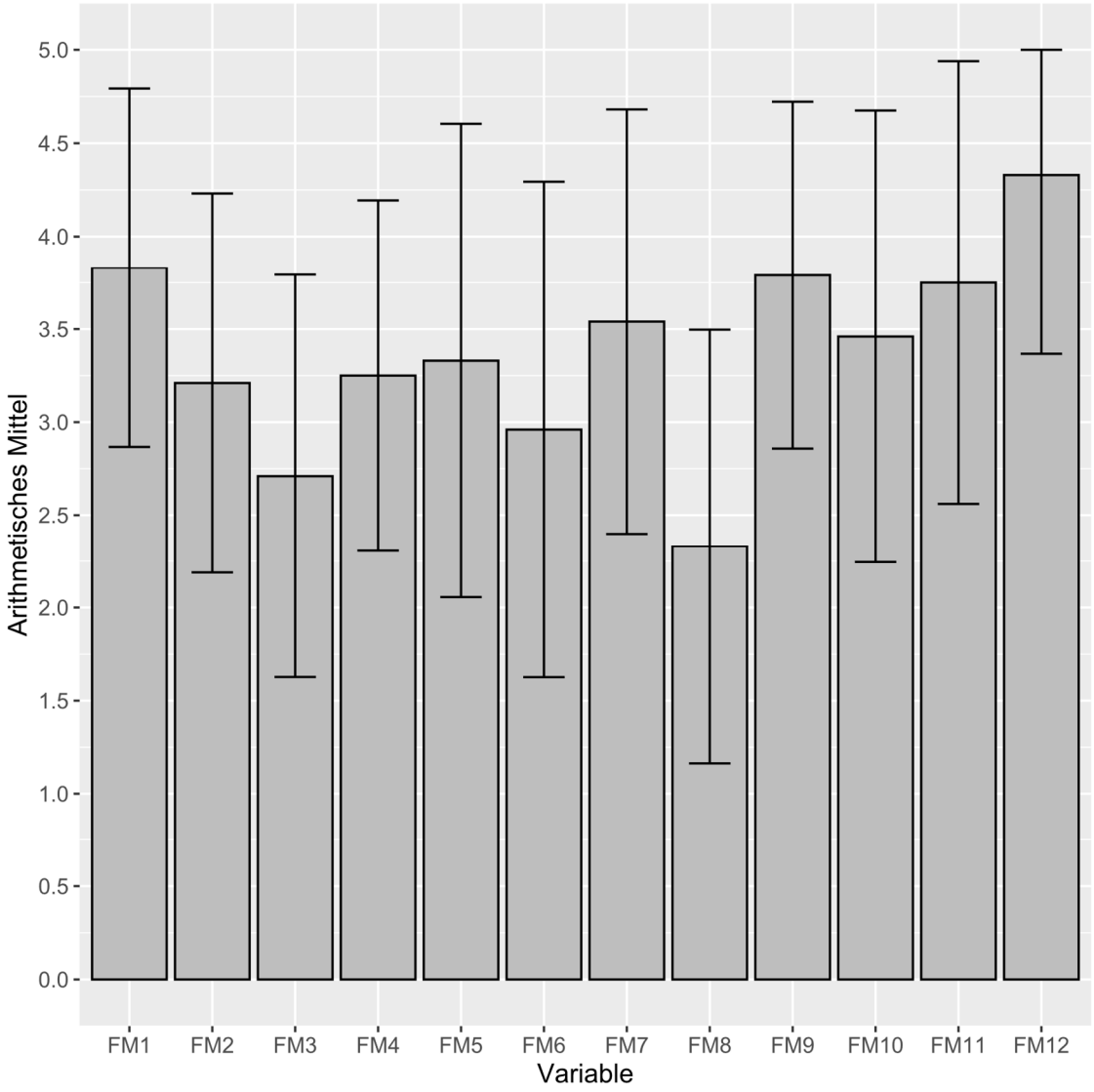


Abbildung 8:Likert-skalierte Wahrnehmung des Spiels durch die Schüler*innen der Experimentalgruppe an der Bismarckschule (BE) [38].

6.3. St. Ursula-Schule

6.3.1. Experimentalgruppe

Der Mittelwert beläuft sich im Pretest der St. Ursula-Schule auf 9,1 Punkte mit einer Standardabweichung von 4,7 Punkten und steigerten sich zum Posttest um einen Punkt auf 10,1 Punkte mit einer leicht gesunkenen Standardabweichung von jetzt nur noch 4,5 Punkten. In Abbildung 9 ist die Häufigkeitsverteilung der Gesamtpunktzahlen über beide Zeitpunkte der Datenerhebung graphisch dargestellt.

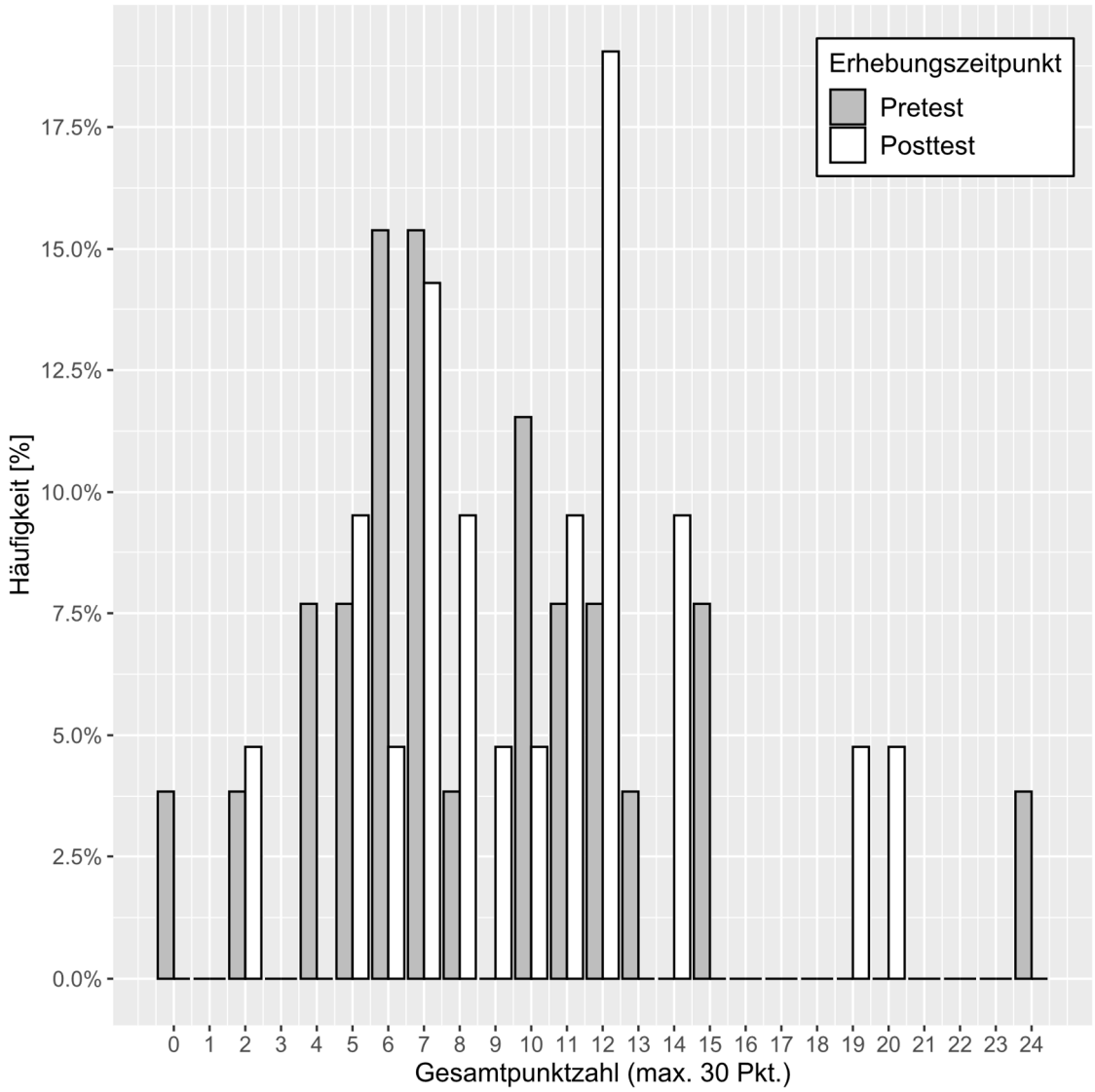


Abbildung 9: Punkteverteilung im Pre- & Posttest der Experimentalgruppe an der St. Ursula-Schule (UE).

Diese lässt eine annähernde Normalverteilung der Messwerte erkennen, die sich vom Pre- zum Posttest jedoch leicht nach rechts in Richtung der niedrigen zweistelligen Punktzahlen verschiebt. Die Streuung der Punktzahlen verringert sich in dieser Gruppe vom Pre- zum Posttest ebenfalls merklich von 24 auf nur noch 18 Punktwerte, die im Leistungsspektrum des Posttests enthalten sind. Wie bereits in den vorherigen Untersuchungen wird hierzu ein gepaarter t-Test mit der Software SPSS durchgeführt. Dieser belegt jedoch, dass es sich bei dem gemessenen Unterschied keinesfalls um einen signifikanten Lernzugewinn handelt ($p = .78$). Die Effektstärke ($d = -.06$) bestätigt zusätzlich, dass der Spieleinsatz keinen messbaren Effekt auf die Leistung der Schüler*innen hatte. Als weiteres Indiz für den potentiellen Einflusses des Spiels auf die Schüler*innen wurde auch

hier eine individualisierte Betrachtung der Entwicklung einzelner Schüler*innen vorgenommen.

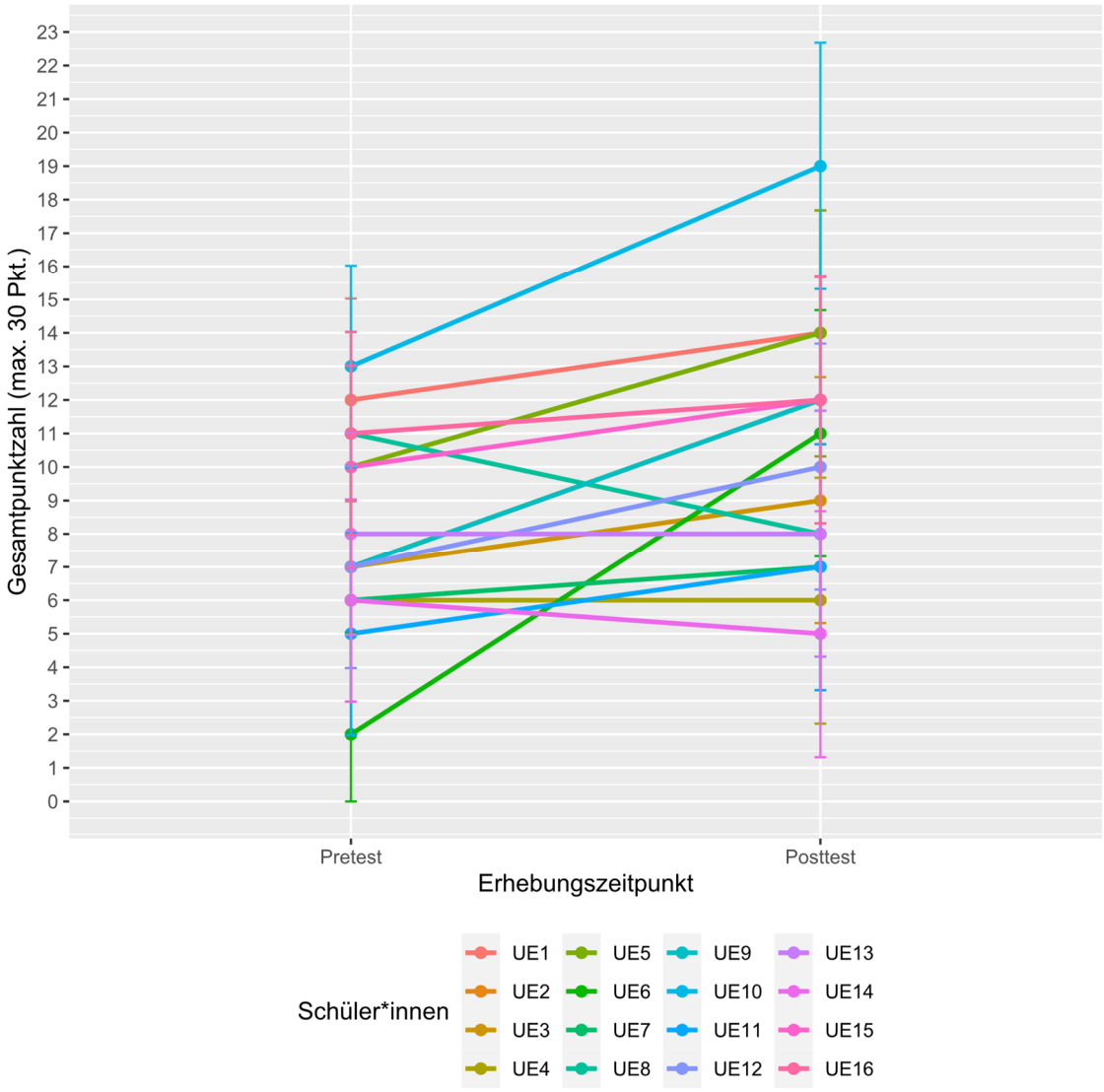


Abbildung 10: Vergleich der Individualleistung am Pre- und Posttest teilnehmender Schüler*innen aus der Versuchsgruppe UE.

Anhand der Teilnehmenden-Codes konnten Schüler*innen identifiziert werden (N = 16) deren Entwicklung in Abbildung 10 graphisch aufgetragen wird (Pretest: S_D = 3,0 Pkt.; Posttest: S_D = 3,8 Pkt.). Bei genauerer Betrachtung fällt auf, dass nur zwei Schüler*innen (UE 7 & UE14) ihr Ergebnis vom Vortest verschlechtern. Bei zwei weiteren (UE4 und UE13) stagniert die Punktzahl, es tritt also weder einer Verschlechterung noch merkliche Verbesserung der Ergebnisse ein. Die verbleibenden 12 Schüler*innen hingegen haben ihre Leistung vom Pre- zum Posttest durchweg verbessert. Wobei mit Ausnahme von einer Person (UE6) sich alle weiteren Veränderungen innerhalb der Fehlerbalken bewegen.

Zum Abschluss wurde auch hier nach einem Kompetenzzuwachs auf Grund des Spieleinsatzes gesucht, der sich durch eine signifikante Veränderung der Häufigkeit an richtigen Lösungen für einzelne Items zeigen würde. Für das Items FA3 ergab sich eine signifikante Reduktion der Häufigkeit an richtigen Antworten ($p \leq .05$) von 46% im Pretest auf nur noch 21% im Posttest. Für das Item FF1 ein hingegen ergab sich eine signifikante Steigerung ($p \leq .01$) der richtigen Antworten von nur 8% im Pretest auf 18% im Posttest.

6.3.2. Kontrollgruppe

Die Schüler*innen der Kontrollgruppe der St. Ursula-Schule erreichten beim ersten Erhebungszeitpunkt ein durchschnittliches Ergebnis von 10,0 Punkten bei einer Standardabweichung von 3,7 Punkten und am zweiten Messzeitpunkt einen geringfügig schlechteren Mittelwert von 9,7 Punkten mit einer erhöhten Standardabweichung von 4,2 Punkten. Aus

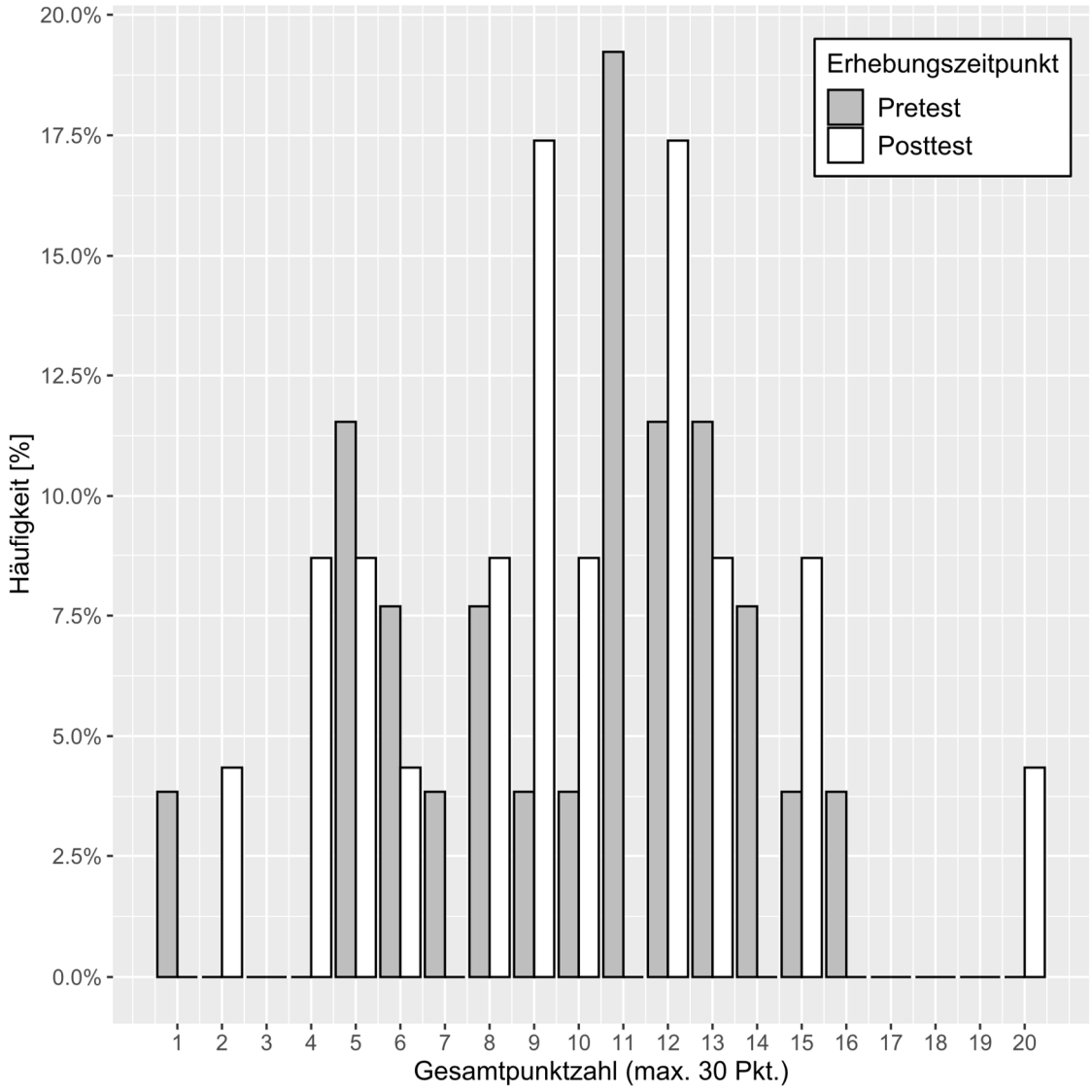


Abbildung 11: Punkteverteilung im Pre- & Posttest der Kontrollgruppe an der St. Ursula-Schule (UK).

der wie in den vorangegangenen Kapiteln dargestellten Häufigkeitsverteilung (s. Abbildung 11) lässt sich im Posttest eine Häufung der Gesamtpunktzahlen um den Mittelwert erkennen. Weiterhin konnte eine Erhöhung der Punktzahlstreuung von 15 Punkten im Pretest auf 18 Punkte im Posttest beobachtet werden. Zur Gewährleistung der Vergleichbarkeit der Experimental- und Kontrollgruppe wurde auch hier die Signifikanz der Mittelwertdifferenz anhand eines t-Tests untersucht. Dessen Ergebnisse belegen weder einen signifikanten Unterschied der Mittelwerte (p = .89), noch einen nachweisbaren Effekt, den der zwischen Pre- und Posttest erteilte Unterricht auf die Leistung im Posttest gehabt haben könnte (d = -.03).

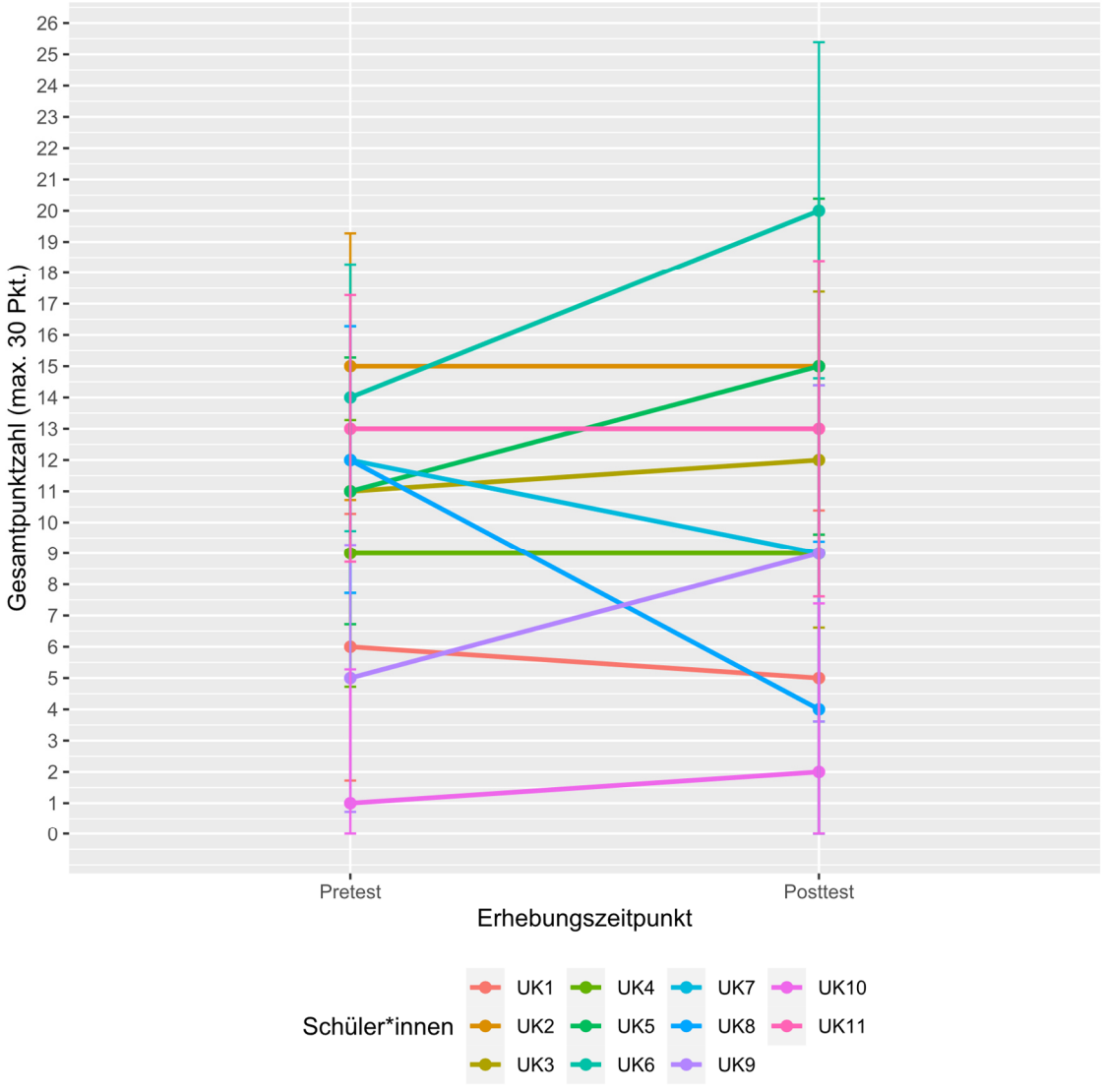


Abbildung 12: Vergleich der Individualleistung am Pre- und Posttest teilnehmender Schüler*innen aus der Versuchsgruppe UK.

Für den individuellen Leistungsvergleich sind in dieser Versuchsgruppe mit Abstand die wenigsten Datensätze zustande gekommen. Lediglich für eine geringe Anzahl (N = 11)

der eigentlich 26 Schüler*innen dieser Klasse konnten Pre- und Posttestergebnisse anhand der Teilnehmenden-Codes einander zugeordnet werden. Die Leistungsentwicklung ist in Abbildung 12 unter Berücksichtigung der Standardabweichungen für den Pre- (SD = 4,3) und Posttest (SD = 5,4) dargestellt. Bei jeweils drei der Schüler*innen ließ sich eine Verschlechterung (UK1, UK7, UK8) oder Stagnation (UK2, UK4, UK11) der Posttestleistung im Vergleich zum Pretest feststellen. Die Leistung der verbleibenden fünf Schüler*innen hat sich auch in dieser Versuchsgruppe vom Pretest hin zum Posttest gesteigert, wenn auch in nicht einem einzigen Fall über das bestehende Fehlerniveau hinaus. Der zur Bewertung von Änderungen in den Häufigkeiten richtiger Antworten für die 30 Testitems durchgeführte t-Test ergab, dass sich in dieser Versuchsgruppe für kein Item signifikante Änderungen zwischen dem Pre- und Posttest nachweisen ließen.

6.3.3. Wahrnehmung des Spiels durch die Schüler*innen

Im Anschluss an die Durchführung des Spiels wurde auch in der Experimentalgruppe der St. Ursula-Schule mittels eines zusätzlichen Fragebogens die intrinsische Motivation der Schüler*innen erhoben. Die Ergebnisse dieser Erhebung sind in Abbildung 13 graphisch aufgetragen. Informationen zur Likert-Skalierung [38] der Messwerte sind Kapitel 6.1.4 zu entnehmen. Anhand der Items FM1 und FM2 lässt sich unschwer feststellen, dass der Einsatz des Spiels die meisten Schüler*innen motiviert hat und sie die Durchführung als unterhaltsam wahrgenommen haben. Die noch hören Zustimmungswerte zu den Items FM11 und FM12 bestätigen die besondere Rolle des Spielens zur Motivation der Schüler*innen und markieren deren ausdrücklichen Wunsch, diese öfter als bisher üblich in die Unterrichtspraxis zu integrieren [14]. Die Daten zeigen also deutlich, dass die Forderung aus der Bildungsforschung, Spiele aus vielerlei Gründen öfter als bisher in den Unterricht einzubinden, von den Schüler*innen mitgetragen wird und die empirisch belegten positiven Auswirkungen des Spielens auch von den Schüler*innen wahrgenommen werden [Vgl. 3, 7, 16, 17, 20, 37]. Leisen benennt als wesentlichen Punkt für gelingendes Lernen die für Schüler wahrnehmbare und offensichtliche Trennung von Lern- und Leistungssituationen, auf die bereits im Kapitel 4.1.2 eingegangen worden ist [40]. Je nach Situation stehen die Schüler*innen variierend unter Druck und passen ihr Verhalten dementsprechend an [40]. Daher wurde der Faktor *„Wahrnehmung von Druck“* [33] von Wilde bereits in der Konstruktion des Erhebungsinstruments berücksichtigt, da dieser die intrinsische Motivation und damit als direkte Folge den Lernerfolg durch die Maßnahme entscheiden beeinflusst [33]. Mit dem Item FM3 ist diese Idee in das Erhebungsinstru-

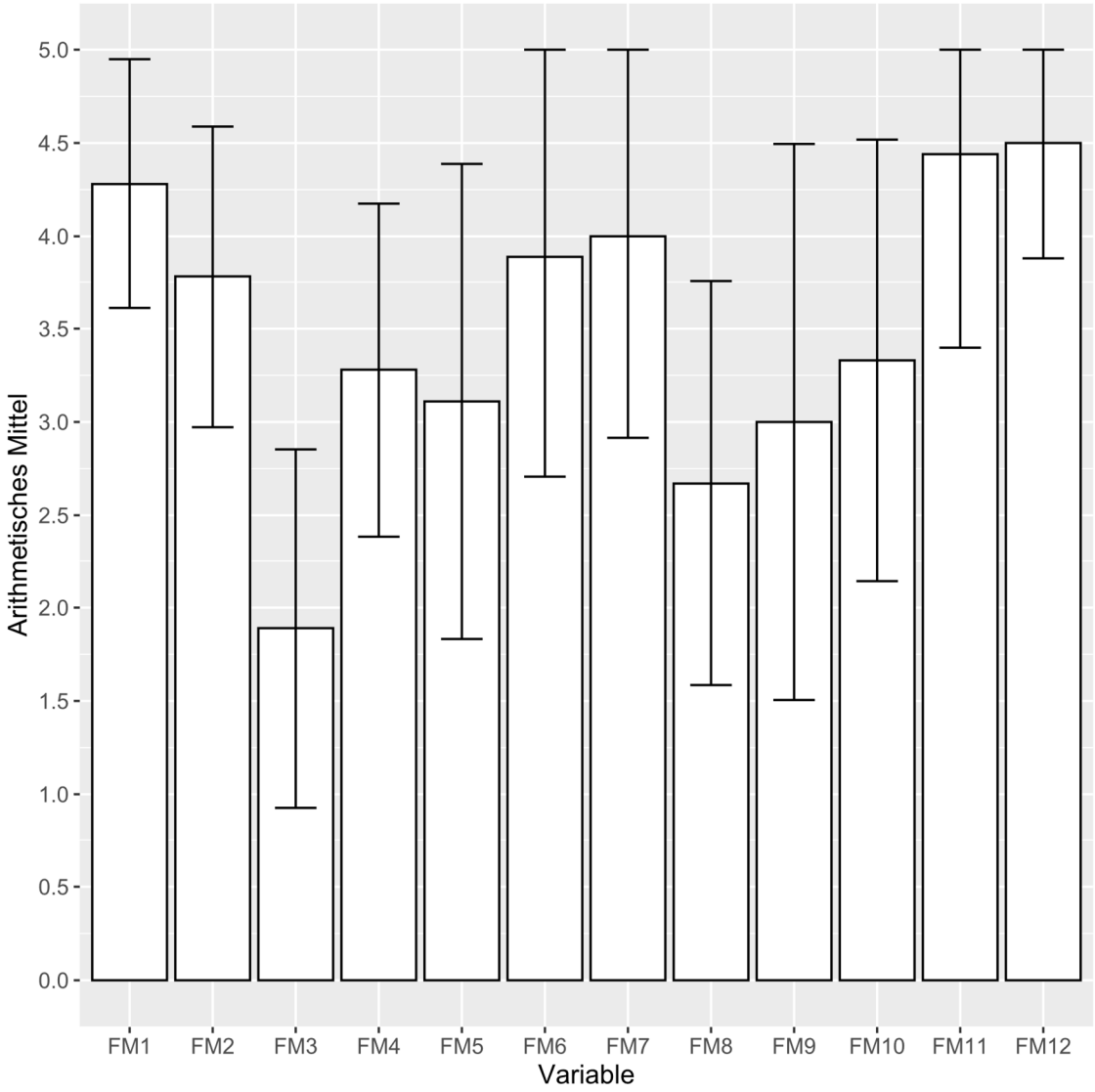


Abbildung 13: Likert-skalierte Wahrnehmung des Spiels durch die Schüler*innen der Experimentalgruppe an der St. Ursula-Schule (UE) [38].

ment übernommen worden. Anhand der niedrigen Zustimmungswerte für dieses Item lässt sich schlussfolgern, dass den Schüler*innen zum einen die Trennung der Phasen des Spielens und der Leistungstests hinreichend deutlich geworden sind und darüber hinaus die Spielsituation von dem Großteil der Schüler*innen als angenehmes Erlebnis empfunden wurde. Daran anschließend lässt sich aus den relativ hohen Werten der Items FM6 und FM7 trotz der nicht signifikanten Ergebnisse bezüglich des Lernzugewinns festhalten, dass die Schüler*innen durchaus das Gefühl hatten, durch das Spiel etwas gelernt zu haben und sie sogar noch stärker der Meinung sind, dass ihnen das Spiel bestehende Wissenslücken aufgezeigt hat. Nichtsdestotrotz wurde von den Spielenden keine nennenswerte Steigerung des Interesses für die Kernphysik berichtet, wie an der Ausprägung des Items FM8 deutlich wird.

7. Interpretation und Diskussion der Ergebnisse

7.1. Beurteilung des Kompetenzzuwachses

7.1.1. Bismarckschule

Werden nun die für sich alleinstehend betrachteten Ergebnisse der Experimentalgruppe und Kontrollgruppe dieser Schule zusammengeführt, so zeigen die Abbildungen 4 und 6, sowie die bestimmten Mittelwerte und Standardabweichungen beider Pretests deutlich, dass sowohl die Kontrollgruppe als auch die Experimentalgruppe mit ungefähr gleichem Vorwissen und unter annähernd identischen Ausgangsbedingungen in die Datenerhebung gestartet ist. Als Erklärung hierfür kann die vergleichbare demographische Zusammensetzung beider Gruppen (s. Kapitel 4.3.2) und auch die Tatsache des bewusst konstant gehaltenen Einflussfaktors Lehrkraft in beiden Klassen herangezogen werden. Bereits in der Vorbereitung der Untersuchung wurde letzterer Aspekt als maßgebliche Bedingung festgesetzt, was sich nun wiederum in der Vergleichbarkeit der Versuchsgruppen als großer Vorteil bemerkbar macht. Diese anfänglich hervorragende Vergleichbarkeit wird jedoch im Posttest nahezu vollständig hinfällig, da es aus organisatorischen Gründen in der Kontrollgruppe zu keinem anderen Zeitpunkt möglich war eine Klassenarbeit zu schreiben. Dies führte zu der unerwarteten Leistungssteigerung der Kontrollgruppe (s. Kapitel 6.2.2) und wirkte sich in Kombination mit dem Ausbleiben eben der in den Hypothesen vorhergesagten Steigerung der Experimentalgruppe als ausschlaggebendes Ereignis für die geringe Aussagekraft der Ergebnisse aus.

Zur Beurteilung und Bestätigung des an den bloßen Rohdaten wahrgenommenen verzerrenden Effekts der Klassenarbeit wurde anhand der Posttestdaten aus der Experimental und Kontrollgruppe ein t-Test für diese voneinander unabhängigen Stichproben durchgeführt, mit welchem der Einfluss des Spieleinsatzes beurteilt werden kann. Dieser bedarf auf Grund der voneinander unabhängigen Stichproben einer anderen Berechnungsgrundlage, die nachfolgend dargestellt wird [59].

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(S_x)^2}{N_x} + \frac{\left(S_y\right)^2}{N_y}}}$$

Formel 5: t-Test für voneinander unabhängige Stichproben [54].

Bemerkung: $\bar{x}$ / $\bar{y}$ = Mittelwert der Versuchsgruppe x/ y; $S_{x,y}$ = Standardabweichung der Versuchsgruppe x/ <y; $N_{x,y}$ = Stichprobengröße der Versuchsgruppe x/ y

Bevor die Ergebnisse beider Versuchsgruppen durch den t-Test miteinander verglichen werden können, musste mittels eines Levene-Tests die Varianzhomogenität beider voneinander unabhängiger Gruppen bestätigt werden [Vgl. 53, 54]. Dieser bestätigte die Varianzhomogenität, sodass die Durchführung des t-Tests aussagekräftige Ergebnisse liefern sollte. Aus diesem t-Test ergab sich jedoch, dass die Nullhypothese nicht verworfen werden kann, da die Mittelwertdifferenz nicht ausreichend signifikant ausfiel ($p = .11$). Damit kann von keinem wesentlichen Zusammenhang zwischen dem Spieleinsatz und einer Leistungssteigerung ausgegangen werden. Die bestimmte mittlere aber negative Effektstärke ($d = -.49$) lässt sich als Indikator für den verzerrenden Effekt der Klassenarbeit verstehen, da das negative Vorzeichen dafür spricht, dass der Spieleinsatz in der Experimentalgruppe, im Vergleich zum regulären Unterricht der Kontrollgruppe, für ein Ausbleiben des Lernzugewinns gesorgt hat. Diese Interpretation ist jedoch vor dem Hintergrund der nicht ausreichenden Signifikanz der Unterschiede ebenfalls kritisch zu betrachten. Die Leistungssteigerung in der Kontrollgruppe widerspricht den in Kapitel 4.1.1 erklärten Regressionseffekten, die für den Posttest eine Häufung der Punkte im mittleren Leistungsspektrum hätten erwarten lassen. Zumindest für die gestiegene Anzahl extrem hoher Punktzahlen können allerdings Sequenzeffekte, also Lerneffekte durch die erneute Verwendung identischer Items im Posttest, als Erklärung herangezogen werden [38]. Die Häufung der besonders niedrigen Punktzahlen im Posttest wird vermutlich auf die gesunkene Motivation zur Durchführung des Leistungstests zum Schuljahresende zurückzuführen sein.

Anhand der Auswertung der an der Bismarckschule erhobenen Daten und dem Vergleich zwischen der Experimental- und Kontrollgruppe lässt sich die Gültigkeit der Hypothese H1 nicht bestätigen. Aus diesem Grund wird von der Gültigkeit der Nullhypothese ausgegangen.

7.1.2. St- Ursula Schule

Werden die Ergebnisse der Experimental- und Kontrollgruppe an der St. Ursula-Schule zusammenhängend betrachtet, spiegelt sich die vergleichbare demographische Zusammensetzung beider Versuchsgruppen (s. Kapitel 4.3.1) in den bestimmten Mittelwerten, Standardabweichungen und diesen Werten zugrundeliegenden Häufigkeitsverteilungen ähnlich zur Bismarckschule wider (s. Abbildungen 9 & 11). Nicht nur die Verteilungsmuster der Diagramme ähneln sich, sondern auch die Mittelwerte und Standardabweichungen stimmen für beide Versuchsgruppen bis auf kleine Unterschiede nahezu überein. Damit wurde die in der Untersuchungsplanung intendierte Vergleichbarkeit der Gruppen,

trotz der nicht zu beeinflussenden Zusammensetzung der Schulklassen, dennoch erreicht. Nichtsdestotrotz sei auch der Einfluss der in beiden Versuchsgruppen identischen Fachlehrkraft auf diese Vergleichbarkeit nicht zu unterschätzen. Die so dicht beieinanderliegenden Ergebnisse der Klassen lassen einen nahezu identisch aufgebauten Unterricht und aus diesem bei den Schüler*innen resultierenden ähnlich hohen Lernerfolg vermuten. Der in Kapitel 6.3.1 beschriebene Zuwachs des Mittelwertes für die Experimentalgruppe lässt sich zusätzlich mit der minimalen Verschlechterung des Mittelwerts in der Kontrollgruppe vergleichen, um die potentielle Wirksamkeit des Spieleinsatzes zu quantifizieren. Hierzu wurden die Posttestergebnisse der Experimental- und Kontrollgruppe einem t-Test für voneinander unabhängigen Versuchsgruppen unterzogen. Der dieser Berechnung vorangegangene Levene-Test attestiert Varianzgleichheit, sodass die Ergebnisse des t-Tests bestätigen können, dass der Spieleinsatz zu keiner signifikanten Steigerung ($p = .76$) des Abschneidens der Experimentalgruppe gegenüber der Kontrollgruppe im Posttest geführt hat. Gemäß geltender Konvention lässt die Beurteilung der Effektstärke ($d = .09$) keinen messbaren Effekt des Spieleinsatzes auf die Ergebnisse des Posttests ableiten. Die Ergebnisse des t-Tests werden durch den Einfluss von Sequenzeffekten unterstützt. Diese können als Erklärung für gestiegene Punktzahlen vom Pre- zum Posttest in der Kontrollgruppe herangezogen werden.

Vor dem Hintergrund der guten Vergleichbarkeit der Experimental- und Kontrollgruppe legt dies den Schluss nahe, dass die Wirkung des Spieleinsatz durch die Nullhypothese beschrieben werden sollte.

7.2. Vergleich beider Schulen

Das Untersuchungsdesign sah neben dem bereits erfolgten Vergleich von Experimental- und Kontrollgruppe zusätzlich eine durch den Einsatz an zwei Schulen erfolgende Parallelisierung vor. Damit können die beschriebenen Ergebnisse an den beiden Schulen nachfolgend miteinander verglichen und in einen breiteren Kontext eingeordnet werden.

Eine wesentliche Voraussetzung für die Vergleichbarkeit der beiden Schulen stellt eine ähnlich repräsentative demographische Zusammensetzung der Untersuchungsgruppen dar. Anhand der im Kapitel 4.3 dargestellten erhobenen demographischen Daten der vier Klassen lässt sich eine nahezu identische Verteilung von ungefähr 50% männlichen und 50% weiblichen Schüler*innen mit nur wenigen Ausnahmen erkennen. Ebenfalls ist die Altersspanne der Schüler*innen in allen Klassen nahezu identisch verteilt. Es gibt nur wenige jüngere oder ältere Schüler*innen, eine deutliche Mehrheit aller Teilnehmenden

war zum Erhebungszeitpunkt 16 Jahre alt und hatte damit das für den 10. Jahrgang typische Alter erreicht, wenn die Einschulung im 6. Lebensjahr erfolgte. Weiterhin lässt sich über das Leistungsspektrum eine nahezu gleiche Verteilung beobachten. Besonders gute und besonders schlechte Noten stellen die Ausnahme dar, die breite Masse gab als letzte Zeugnisnote im Fach Physik eine befriedigende bis gute Leistung an.

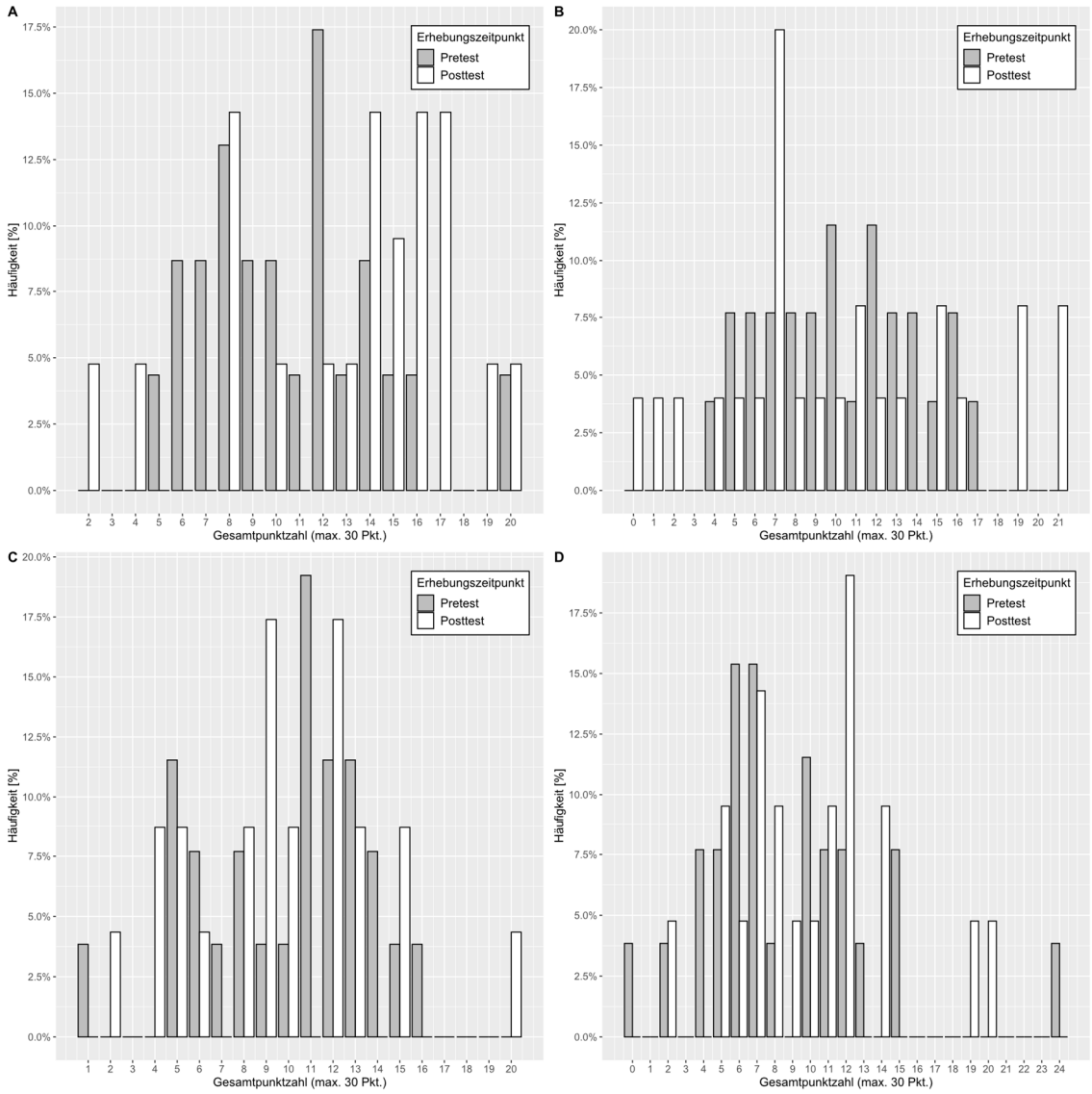


Abbildung 14: Übersicht der Punkteverteilung der Pre- & Posttests aller vier Untersuchungsgruppen.
Bemerkung: **A** Bismarckschule Kontrollgruppe (BK), **B** Bismarckschule Experimentalgruppe (BE), **C** St. Ursula-Schule Kontrollgruppe (UK), **D** St. Ursula-Schule Experimentalgruppe (UE).

Zuletzt erhoben wurde die Vorerfahrung der Schüler*innen mit Escape-Räumen oder derartigen Spielen, wobei sich auch hier eine ungefähre Gleichverteilung über alle Versuchs-

gruppen einstelle, da die deutliche Mehrheit angab, noch nie in Kontakt mit dieser speziellen Spielkategorie gekommen zu sein. Zur Vergleichbarkeit lässt sich also resümierend festhalten, dass trotz der fehlenden Möglichkeit der Randomisierung der Versuchsgruppen, mit der Auswahl der vier Klassen eine ausreichend vergleichbare Versuchspopulation zusammengestellt wurde.

Im Anschluss an die Feststellung der Vergleichbarkeit auch über die einzelnen Schulen hinweg lässt sich der Leistungsvergleich in bekannter Schrittweise auf die Gemeinsamkeiten und Unterschiede beider Schulen wiederholen. Um einen ersten Eindruck über alle Versuchsgruppen hinweg zu erhalten werden in Abbildung 14 die Häufigkeitsverteilungen der Gesamtpunktzahlen der vier Untersuchungsgruppen gemeinschaftlich dargestellt. Der Vergleich, der sich aus diesen Verteilungen ergebenden Mittelwerte lässt mit einem Wert der nur leicht um 10 Punkte pendelt, ein nahezu identisches Leistungsniveau über alle Gruppen hinweg vermuten. Auffällig hingegen ist die Tatsache, dass sich hohe zweistellige Punktzahlen, größer 15 Punkte, an der Bismarckschule häufiger finden, als an der St. Ursula-Schule. Zu deren Einordnung muss jedoch der Effekt der Klassenarbeit in der Kontrollgruppe der Bismarckschule mit betrachtet werden. Besonders im Posttest dieser Gruppe ist eine deutliche Zunahme im angegebene Punktebereich zwischen Pre- und Posttest zu erkennen. Werden aufgrund dieses Effekts die Kontrollgruppen außen vor gelassen und wird sich auf die Experimentalgruppen beider Schulen konzentriert, so ist an der St. Ursula-Schule vom Pre- zum Posttest eine leichte Verschiebung der Verteilung hin zu höheren Punktzahlen zu erkennen, wohingegen die Punkteverteilung an der Bismarckschule sowohl hin zu höheren als auch niedrigeren Punktzahlen divergiert, wobei die mittleren Punktzahlen dann im Posttest seltener als noch im Pretest auftreten.

Unter Zuhilfenahme der Abbildung 15 lassen sich die individuellen Leistungsentwicklungen der Schüler*innen beider Schulen miteinander vergleichen. In den Kontrollgruppen spiegelt sich erneut der Klassenarbeitseffekt in einer Stagnation des mittleren Leistungsstandes an der St. Ursula-Schule und einem deutlichen Leistungszuwachs an der Bismarckschule wider. Wird auch hier die Experimentalgruppen herangezogen, dreht sich die Beobachtung um. An der St. Ursula-Schule kam es bis auf wenige Ausnahmen zu einem Anstieg der Punktzahlen vom Pre- zum Posttest, währenddessen sich die Punktzahlen an der Bismarckschule eher verschlechterten. Für die Vergleichbarkeit und Aussagekraft der Ergebnisse muss hier jedoch berücksichtigt werden, dass aufgrund von nicht zuzuordnenden Teilnehmenden-Codes große Teile der erhobenen Daten nicht berück-

sichtigt werden konnten und die Repräsentativität der Daten durch die verringerte Stichprobengröße leidet. Vermutlich lassen sich aus diesem Grund keine weiteren wesentlichen Tendenzen in den Daten erkennen.

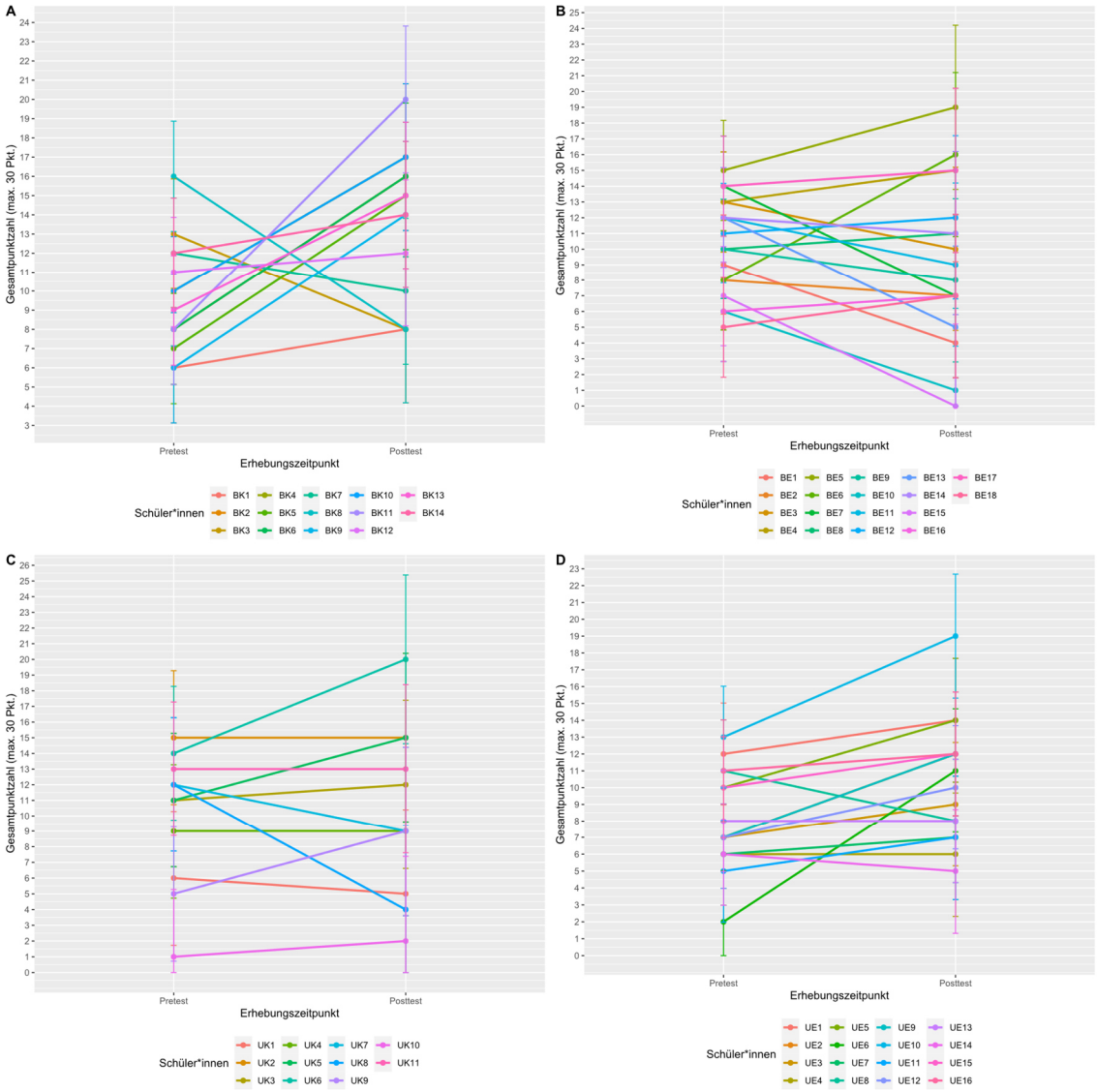


Abbildung 15: Übersicht des Leistungsvergleichs im Pre- & Posttest aller vier Versuchsgruppen.
Bemerkung: **A** Bismarckschule Kontrollgruppe (BK), **B** Bismarckschule Experimentalgruppe (BE), **C** St. Ursula-Schule Kontrollgruppe (UK), **D** St. Ursula-Schule Experimentalgruppe (UE).

Neben der Leistungsmessung war ein weiteres Anliegen dieser Arbeit die Wahrnehmung der des Spiels durch die Schüler*innen zu erheben und daraus Rückschlüsse für dessen zukünftigen Einsatz im Unterricht zu schließen. Abbildung 16 führt die erhobenen Daten aus den Experimentalgruppen beider Schulen zusammen, sodass auch anhand dieser ein

Vergleich beider Schulen erfolgen kann. Grundsätzlich ist festzustellen, dass die Wahrnehmungen des Spieleinsatzes an der St. Ursula-Schule stärker zu Extremwerten tendiert als an der Bismarckschule. Diesbezüglich ist nicht nur eine stärkere Zustimmung, sondern auch Ablehnung bestimmter Items als an der Bismarckschule festzustellen.

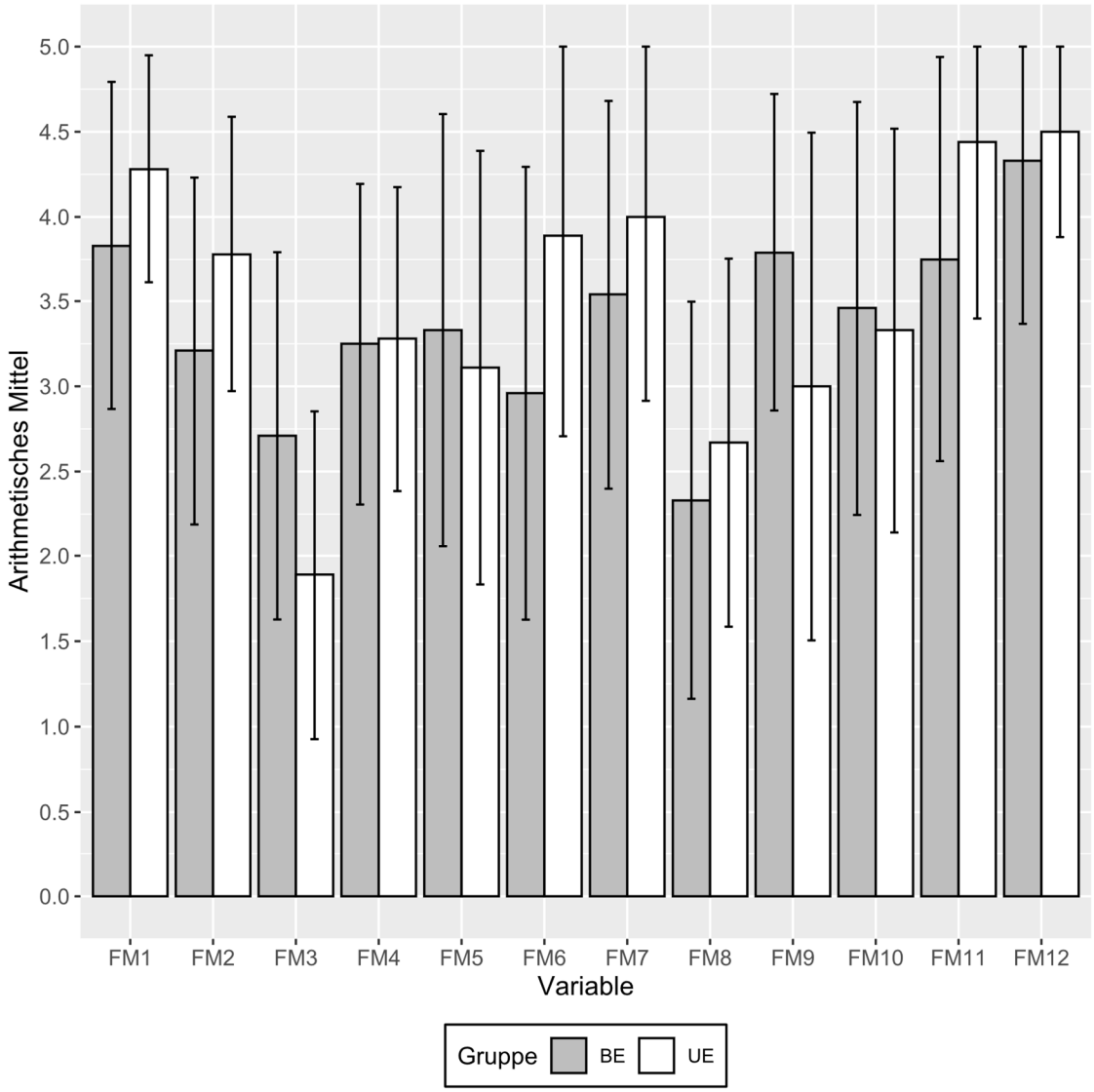


Abbildung 16: Vergleich der Likert-skalierten Wahrnehmung des Spiels durch die Schüler*innen der Experimentalgruppe beider Schulen. Bismarckschule Experimentalgruppe (BE); St. Ursula-Schule Experimentalgruppe (UE) [38].

Anhand der Fehlerbalken in Abbildung 16 wird jedoch auch ersichtlich, dass es sich bei diesen keineswegs um repräsentative Unterschiede zwischen der Wahrnehmung des Spieleinsatzes an beiden Schulen handelt. Dennoch lässt sich die motivierende Eigenschaft und der spaßbringende Charakter des Spieleinsatzes über beide Versuchsgruppen hinweg anhand mehrerer Items (FM1, FM2, FM11, FM12) bestätigen. Zusätzlich wurde

die Spieldurchführung über beide Gruppen hinweg nicht als Drucksituation wahrgenommen, wie sich an der niedrigen Ausprägung des Items FM3 sehen lässt. Der wahrgenommene Kompetenzzuwachs durch das Spiel fiel jedoch an der St. Ursula-Schule stärker aus, als an der Bismarckschule (FM6, FM7). Entgegengesetzt zu dieser Beobachtung wurde die Gruppengröße (FM9), die in der Literatur als maßgeblicher Faktor für den Lerneffekt dargestellt wird, an der Bismarckschule als angemessener im Vergleich zur St. Ursula-Schule wahrgenommen [18]. Erklärt werden kann dies durch die vielseitig in beiden Experimentalgruppen beobachteten Nebenbeschäftigungen der Schüler*innen während der Spielphasen. Die überdimensionierte Größe der Spielgruppen hat dazu geführt, dass von den zwischen 10 und 12 Spielenden lediglich 4-5 Schüler*innen aktiv an den Rätseln gearbeitet haben und die Verbleibenden die Unterrichtszeit zu ihrer eigenen Zufriedenheit gestaltet haben. Dies führte schlussendlich zu der überdurchschnittlich positiv eingeschätzten Gruppengröße, da ein Großteil der Schüler*innen die Stunde zu ihrer freien Verfügung und nicht als Lernsituation wahrgenommen haben.

Zusammenfassend lässt sich der Vergleich beider Schulen mit der Feststellung abschließen, dass nicht nur die Experimental- mit ihren dazugehörigen Kontrollgruppen an den beiden Schule angemessen miteinander vergleichbar sind, sondern dass die Versuchsgruppen auch über die jeweilige Schule hinaus, auf Grund ähnlicher demographischer Zusammensetzung repräsentative Ergebnisse geliefert haben. Für diese Schlussfolgerung spricht auch das ähnliche Leistungsspektrum aller Klassen, welches durch eine vergleichbare Unterrichtsgestaltung und gut auf das Kerncurriculum abgestimmte Unterrichtsinhalte beider Fachlehrer erreicht worden ist.

8. Einschränkungen der Ergebnisse

Im Laufe des Erhebungsprozesses haben sich durch Beobachtungen während der Spiel- und Testphasen, in Gesprächen mit den unterrichtenden Lehrkräften und auch in der Auswertung der Daten verschiedenste Einflüsse aufgetan, die die Aussagekraft der Untersuchungsergebnisse beeinflussen könnten oder beeinflusst haben. Einige von diesen in der Vorbereitung der Untersuchung unvorhersehbaren Einflüsse wurden in vorangegangenen Abschnitten dieser Arbeit bereits kurz erwähnt und sollen nun ausführlicher beleuchtet werden, damit unter Berücksichtig dieser Hemmnisse eine abschließende Bewertung der Ergebnisse vorgenommen werden kann.

Es liegt nahe, diese Auflistung mit den Einflüssen zu beginnen, die bereits vor dem eigentlichen Erhebungszeitpunkt eine Einschränkung der Daten vorherbestimmt haben. Gemeint sind damit spezielle Eigenschaften, die das gewählte Forschungsdesign mit sich bringt. Hier sei besonders die Anzahl der Testzeitpunkte hervorzuheben, da laut Döring bereits eine Erhöhung von zwei auf drei Testzeitpunkte die Reliabilität der zu messenden Veränderungsmaße signifikant erhöht [38]. Damit könnten die Auswirkungen eines nicht ausreichend reliablen Messinstruments ausgeglichen werden. Jedoch wurde in Kapitel 5.3.2 eine zufriedenstellende Reliabilität bestimmt, sodass diese Maßnahme auch aufgrund des potentiell negativen Effekts eines weiteren Leistungstests auf die Motivation der Schüler*innen, vermutlich keinen wesentlichen Effekt hätte haben können. In Anbetracht des sowieso eng getakteten Zeitplans der Datenerhebung und den terminlichen Einschränkungen die durch Feiertage und Ferien entstanden sind, war ein weiterer Erhebungszeitpunkt praktisch nicht umsetzbar. Darüber hinaus wurden die einzelnen Frageitems für den Pre- und Posttest, genauso wie die Antwortmöglichkeiten bewusst identisch gehalten, um die Vergleichbarkeit der gemessenen Leistung zu gewährleisten. Die daraus resultierenden Sequenzeffekte zeigen sich in den Posttestdaten durch verzerrte Messwerte, wie es beispielsweise in der Kontrollgruppe der Bismarckschule zu sehen ist. Weiterhin sei an dieser Stelle auch die nicht zufriedenstellende Konstruktvalidität des Messinstruments erwähnt. Die durchgeführte explorative Faktorenanalyse hat die, in den theoretischen Vorbetrachtungen entwickelten, sieben inhaltlichen Teilbereiche des Leistungstests nicht bestätigt. Stattdessen ergab sich nur für den Fall eines einzelnen verbliebenden Faktors ein signifikantes Ergebnis. Dies legt die Interpretation nahe, dass es sich bei besagtem Faktor um das Fachwissen zum Themenbereichs Kernphysik handeln kann, da alle Items stark mit diesem Faktor korrelieren. Das wiederum eröffnet die Frage, ob

eine weitere Unterscheidung in Subkategorien dieses Fachgebiets überhaupt empirisch mittels Faktorenanalysen nachgewiesen werden könnte und damit sinnvoll ist.
Im Anschluss an den Einflussfaktor Erhebungsinstrument können nun die verschiedenen Faktoren während der Datenerhebung genauer beleuchtet werden. Einleitend sei hier das in Kapitel 4.1.2 angesprochene Problem der nicht identischen Erhebungszeiträume erläutert. Vor Beginn der Datenerhebung war die Parallelisierung der zu untersuchenden Gruppen geplant worden, um die Vergleichbarkeit der Daten zu gewährleisten. Auch wenn in den bisherigen Erläuterungen deutlich geworden ist, dass die Ergebnisse aller Gruppen einander in guter Näherung entsprechen, hat die Terminverschiebung und damit verbundene Verlängerung des Abstandes zwischen Pre- und Posttest an der Bismarckschule zumindest für einen längeren Zeitraum, in dem die Schüler*innen etwas hätten lernen können, gesorgt. In Kombination mit der eingeschobenen Klassenarbeit in der Kontrollgruppe hat sich diese Zeitdifferenz in Form der erhöhten Punktzahlen wahrnehmbar, aber nicht signifikant geäußert. Überdies hatte auch die immer noch anhaltende Corona-Pandemie einen nicht zu unterschätzenden Einfluss auf die erhobenen Daten. Durch krankheitsbedingt fehlenden Schüler*innen ergab sich eine starke Fluktuation innerhalb der Versuchsgruppen. Die Schnittmenge der Schüler*innen die sowohl Pre- und Posttest als auch die Spielphasen durchlaufen hatte, reduzierte sich auf zum Teil weniger als zehn Personen. Beschriebene Fluktuation brachte gleich mehrere Probleme mit sich, die die Aussagekraft der erhobenen Daten reduziert. Zuerst sei die allgemeine Reduktion der Stichprobengröße und die darunter leidende Signifikanz der Ergebnisse als Folge anzuführen. Geeignete Größen, die für die Aussagekraft dieser Untersuchung herangezogen wurden (Testgüte, Effektstärke, Signifikanz von Korrelationen, etc.), profitieren von möglichst großen Stichproben. Weiterhin bewirkte der krankheitsbedingte Ausfall von Schüler*innen einen Ausfall der Lernmöglichkeit. Das entwickelte Messinstrument kann keinen Effekt des Spieleinsatzes zeigen, wenn die teilnehmenden Schüler*innen krankheitsbedingt bei der Spieldurchführung fehlen. Besonders für die Experimentalgruppe der Bismarckschule könnte dies eine Erklärung für die niedrigen Leistungswerte im Posttest liefern, da auf Grund einer Klassenfahrt zwischen Pre- und Posttest eine besonders große Fluktuation unter den Schüler*innen herrschte. Als wohl letzte Folge der Pandemie sei das krankheitsbedingte Fehlen des Versuchsleiters während der Pretests beide Versuchsgruppen an der St. Ursula-Schule zu benennen. Der daraus resultierende Kontrollverlust über die Datenerhebung und das vollständige und recht kurzfristige Übertragen dieser

Aufgabe auf die betreuende Lehrkraft haben maßgeblich zur Verringerung der Aussagekraft dieser zwei Datensätze beigetragen. Bemerkbar machte sich dies in besonderer Weise in den Teilnehmenden-Codes, die zum anonymisierten Vergleich der individuellen Entwicklung vom Pre- zum Posttest benötigt wurden. Allein während der Erhebung dieser zwei Datensätze (UVE & UVK) kam es zu einer Häufung von elf nicht dem verlangten Muster entsprechenden oder schlicht doppelten Teilnehmenden-Codes, die für die Auswertung nicht mehr zur Verfügung standen und somit die aus der Stichprobengrößenreduktion resultierenden Effekte verursachten. In den sechs weiteren Terminen der Datenerhebungen, die unter Anwesenheit der Versuchsleitung durchgeführt wurden, kam es zu nur einem einzigen Fall von fehlerhaften Teilnehmenden-Codes, sodass dieser Effekt eindeutig auf das krankheitsbedingte Fehlen der Versuchsleitung zurückzuführen ist. Eine zeitliche Verschiebung der Erhebungstermine bis nach der Genesung wäre auf Grund der zeitlichen Nähe zu den Sommerferien nicht möglich gewesen, sodass die aus der Abwesenheit resultierenden Schwierigkeiten in Kauf genommen werden mussten.

In der vorangegangenen Erläuterung wurde bereits deutlich, dass auch der Vergleich von individuellen Entwicklungen der Schüler*innen mit gewissen Problemen verbunden war, die vor allem aus einer sehr geringen, teilweise nur knapp zweistelligen, Stichprobengrößen resultierten. Weiterführend haben auch die Schwierigkeiten mit den Teilnehmenden-Codes zum Ausschluss einiger Datensätze geführt, was die Stichprobengröße immer weiter reduzierte. Zumindest letzteres Problem ließe sich in zukünftigen Untersuchungen dieser Art durch einen Ausschluss des Faktors Schüler*innen in der Konstruktion der Teilnehmende-Codes verhindern. Die Zuordnung von vollständig randomisierten Hash-Codes durch ein entsprechendes Computerprogramm beispielsweise würde in der Vorbereitung der Datenerhebung einen Mehraufwand bedeuten, der an dieser Stelle aber zu einer maßgeblichen Verbesserung der Ergebnisse führen kann.

Insbesondere die Größe der Spielgruppen hat zu Recht im Laufe dieser Arbeit an den verschiedensten Stellen immer wieder Erwähnung gefunden. Demzufolge ist der Gruppengröße auch in diesem Kapitel eine gewissen Rolle zugesprochen worden. Mehrfach wurde erwähnt, dass die einschlägige Literatur eine Gruppengröße von 4-6 Schüler*innen für einen effektiven Einsatz von Exit-Spielen vorsieht [Vgl. 18, 24, 26,27]. Bei Klassengrößen zwischen 20 und 30 Schüler*innen und einer Anzahl von lediglich zwei Spielexemplaren konnte die ideale Gruppengröße selbstverständlich nicht eingehalten werden. Die resultierenden Gruppengrößen von deutlich über zehn Schüle*innen führten dazu, dass pro Gruppe lediglich 4-6 Schüler*innen konzentriert an den Rätseln arbeiteten und

der Rest sich mit unterrichtsfernen Inhalten beschäftigte. Ganz im Sinne einer bereits zu Beginn dieses Kapitels getroffenen Aussage lässt sich auch hier formulieren, dass ein Treatment-Effekt nur gemessen werden kann, wenn die Versuchspersonen tatsächlich ein Treatment durchlaufen haben. Die wesentliche Folge der zu geringen Anzahl an Spielexemplaren war, dass sich mindestens die Hälfte der beiden Experimentalgruppen über beide Spielwochen wenig bis überhaupt nicht mit dem Spiel auseinandergesetzt haben und damit auch keine Chance hatten, von dem Spieleinsatz zu profitieren und ihre bisher erworbenen inhalts- und prozessbezogenen Kompetenzen zu wiederholen, zu festigen und zu transferieren, um dadurch im Posttest eine gewisse Leistungssteigerung erreichen zu können.

Als weiteren Aspekt soll noch die Motivation der Schüler*innen angeführt werden. Ungeachtet der gemessenen hohen Motivation durch den Spieleinsatz in den Experimentalgruppen war über alle vier Versuchsgruppen in den Testphasen ein merklicher Motivationsverlust zu verspüren, der sicherlich auf das endende Schuljahr und die nahenden Sommerferien zurückzuführen ist. Ein negativer Einfluss dieser fehlenden Motivation auf die Testergebnisse lässt sich nicht empirisch belegen, jedoch anhand von Schüler*innenäußerungen während und nach der Erhebung dennoch vermuten. Bereits in der Entwicklung des Erhebungsinstruments (s. Kapitel 5.2) wurde mit der Option „keine Antwort“ eine Methode zur Vermeidung von verzerrenden Effekten des Ratens von Lösungen in den Leistungstest integriert. Die Motivation einiger Schüler*innen schien zum Schuljahresende bereits soweit abgesunken zu sein, dass es nicht mehr zum Lesen des Einleitungstextes des Tests ausreichte, wie durch die Aussage *„Wenn man geraten hat ging der Test viel schneller!“* verdeutlicht wird.

Den letzten Aspekt dieses Kapitels stellt die Beobachtung dar, dass die anfängliche Motivation der Schüler*innen der Experimentalgruppen in der zweiten Unterrichtsstunde der Spielphase merklich zurückgegangen war. Erklären lässt sich diese Beobachtung anhand der bereits erwähnten Flow-Theorie [11]. Die für das spielerische Lernen existentielle Immersion und das Flow-Erleben der Schüler*innen wurde nach der ersten Unterrichtsstunde unterbrochen und konnte von diesen in der fortführenden Unterrichtsstunde nach einer zeitlichen Unterbrechung von einer Woche nur schwer wieder erreicht werden, was sich für Beobachtende sichtbar auf die Motivation zur Beendigung des Spiels auswirkte. Darüber hinaus sei an dieser Stelle jedoch anzumerken, dass die in der Planung angesetzte Durchführung des Spiels über zwei Unterrichtsdoppelstunden (s. Kapitel 5.1) nur bei der Hälfte aller Spielgruppen erfolgreich verlaufen ist. Sowohl an der Bismarck- als auch an

der St. Ursula-Schule hat es in den Experimentalgruppen jeweils die Hälfte der Klassen geschafft, das Spiel in der vorgegebenen Zeit zu bearbeiten. Für die andere Hälfte der Klassen musste das Spiel mangels Zeit abgebrochen werden. Daraus resultierte wiederum für diese Hälfte der Klasse die fehlende Möglichkeit, vollständig vom Spieleinsatz zu profitieren, was sich ebenfalls auf die Posttestdaten auswirkte. Diese Problematik war auch Veldkamp in einer Untersuchung aufgefallen und konnte bis heute noch keiner adäquaten Lösung zugeführt werden [21]. Denkbar wäre jedoch eine weitere Reduktion der Spielinhalte, sodass dieses gekürzte Spiel in einer Unterrichtsdoppelstunde von 90 Minuten vollständig abgeschlossen werden könnte. Eine mögliche Reduktion könnte lediglich aus den ersten fünf Rätseln dieses EXIT-Spiels bestehen, da alle Versuchsgruppen diese erfolgreich in 90 Minuten absolviert hatten. Thematisch würde das Spiel damit lediglich den Aufbau des Atoms, sowie die Eigenschaften der α-, β- und γ-Strahlung sowie die Karlsruher Nuklidkarte behandeln.

Bezüglich der Aussagekraft der Ergebnisse dieser Erhebung lässt sich also festhalten, dass relevante Ergebnisse geliefert werden konnten, für zukünftige Untersuchungen jedoch weiterhin Verbesserungspotential besteht. Nichtsdestotrotz lieferte die Arbeit wichtige Erkenntnisse, die den zukünftigen Einsatz nicht nur dieses speziellen Exit-Spiels beeinflussen sollten und werden.

9. Zusammenfassung und Ausblick

Motiviert wurde diese Untersuchung über den bisher nur wenig erforschten Einsatz und damit verbundenen Nutzen von Escape-Räumen und ähnlichen Spielen im Schulunterricht aus der sich die Forschungsfrage FF1 ergeben hat [11]. Aus den in Kapitel 6 dargestellten Ergebnisse der Leistungstests aller vier Versuchsgruppen lässt sich schlussfolgern, dass sich auf Grund des Spieleinsatzes nicht der in Hypothese H1 formulierte Effekt eingestellt hat, und dass damit die Nullhypothese H0 als zutreffend angenommen werden muss. Maßgeblich für diese Schlussfolgerung verantwortlich sind die nicht signifikanten Differenzmaße der Leistungstests beider Experimentalgruppen die im Zusammenspiel mit dem Vergleich zu den Kontrollgruppen nahelegen, dass der vermutete positive Effekt des Spiels nicht vorhanden ist, oder unter den gegebenen Bedingungen nicht messbar ist. Auch wenn die Hypothese H1 nicht bestätigt werde konnte, bedeutet dies nicht, dass der Spieleinsatz keinen Effekt auf die Schüler*innen gehabt hat. Anhand der vorangegangen erläuterten einschränkenden Bedingungen wird deutlich, dass es eine Vielzahl von Einflüssen gab, die die Aussagekraft der Ergebnisse beeinflusst haben. Es ist daher möglich, dass diese Effekte die Daten dahingehend beeinflusst haben und ein vermuteter Lerneffekt verschleiert worden ist. Für weiterführende Untersuchungen wäre es daher ratsam, die angesprochenen Komplikationen anhand der Verbesserungsvorschläge, wie beispielsweise der Hash-Codierung, zu umgehen und so vielleicht auch einen kleinen Effekt des Spieleinsatzes signifikant messbar zu gestalten, wie es bereits bei anderen Untersuchungen gelungen ist [29]. Darüber hinaus wurden mittels des Feedbackfragebogens die intrinsische Motivation der Schüler*innen zur Beantwortung der Forschungsfragen FF2 und FF3 gemessen. Die vermutete Hypothese H2 bezüglich der hohen Motivation der Schüler*innen durch den Spieleinsatz (FF2) konnte anhand der erhobenen Daten aus beiden Experimentalgruppen bestätigt werden und spiegelt damit die Ergebnisse der aktuellen Forschung wider [18, 19, 24, 25, 34]. Bezüglich der Hypothese H3 lässt sich aus den Daten kein aussagekräftiger Trend erkennen, sodass davon auszugehen ist, dass das Spiel zwar die Motivation der Schüler*innen, jedoch nicht deren Interesse für die Kernphysik beeinflusst hat.

Aus der statistischen Auswertung der Daten lassen sich ebenfalls Rückschlüsse bezüglich des entwickelten Erhebungsinstruments ziehen. Vormerklich sei hier die nicht ausreichend nachweisbare Konstruktvalidität genannt. Die bereits ausführlich beleuchtete explorative Faktorenanalyse hat das dem Erhebungsinstrument zugrundeliegende Konstrukt

mit den dazugehörigen Themenbereichen und Lernzielen nicht bestätigen können. Neben der getroffenen Schlussfolgerung einer nicht ausreichenden Konstruktvalidität, die sich in den Daten widerspiegelt, sollte für zukünftige Untersuchungen jedoch ein weiterer Aspekt berücksichtigt werden. Der aktuelle Stand der Forschung lässt keine Aussagen darüber treffen, ob und inwieweit es möglich ist, den für die Schule sehr eng gefassten Themenbereich Kernphysik überhaupt in weitere Subkategorien zu unterteilen, die dann noch statistisch signifikant nachgewiesen werden können. Diesen Eindruck bestätigen Daten aus der Faktorenanalyse, in welchen, unter der Bedingung einer genügenden Signifikanz, nahezu alle Items auf nur einen Faktor korrelieren. Dies legt eine Interpretation dieses einen Faktors als Fachwissen zur Kernphysik nahe und damit verbunden die Hypothese, dass keine weitere Unterscheidung in Teilkompetenzen zu diesem Fachgebiet statistisch signifikant nachgewiesen werden können. Eine anschließende Untersuchung besagter Problematik in weiterführenden Arbeiten könnte helfen, die Aussagekraft der Ergebnisse dieser Erhebung angemessener einordnen zu können.

Neben den Ergebnissen der statistischen Auswertung der in den Leistungstests und abschließenden Umfrage erhobenen Daten konnten auch eine Reihe von Beobachtungen über alle Versuchsgruppen hinweg getätigt werden, die sich für nachfolgende Projekte und Datenerhebungen als hilfreich erweisen könnten. Hierbei soll im Wesentlichen die bereits mehrfach erwähnte Problematik der Gruppengröße in den Spielphasen, als markanten Einfluss auf den möglichen Lerneffekt des Spiels, betrachtet werden [18, 24, 26, 27]. Zu Beginn der Datenerhebung lag das zu untersuchende Spiel in nur zwei gedruckten Exemplaren vor, da in der Entwicklungsphase fälschlicherweise eine Gruppengröße von knapp über zehn Schüler*innen als adäquat angenommen worden ist [9]. Diese Annahme konnte anhand einschlägiger Untersuchungen [18, 24, 36, 27] bestätigt werden und spiegelte sich nachfolgend in den Spielphasen wider. Daher lässt sich schlussfolgern, dass für den zukünftigen Einsatz des Spiels im Unterricht weitere Exemplare erstellt werden müssen. Empfehlenswert wäre eine Anzahl von mindestens sechs Spielexemplaren, da so bei üblichen Klassengrößen von über 20 bis hin zu knapp über 30 Schüler*innen eine Gruppengröße während der Spieldurchführung von 4-6 Schüler*innen gewährleistet werden kann. Diese Gruppengröße wird neben der Literatur zu konkreten Untersuchungen zu ESCAPE-Räumen und EXIT-Spielen auch allgemein für kooperative Lernphasen als angemessene und lernförderliche Gruppengröße beschrieben [6, 60]. Neben dem positiven Effekt auf die Schüler*innen könnte durch die weitere Produktion von Spielexemplaren ein

ähnliches Konzept zu dem bereits existierenden Radlab eingeführt werden [61]. Lehrer*innen könnten sich so vollständige Klassensätze und alle notwendigen Materialen des Spiels am Institut für Radioökologie und Strahlenschutz ausleihen. Weiterhin erhalten diese vom Institut fachkundige Unterstützung bei der Durchführung und können ohne großen Aufwand nach der Verwendung das Material wieder zurückgeben, sodass es für den nächsten Einsatz vorbereitet werden kann [61]. Damit würde der mit dem Spieleinsatz verbundene Vorbereitungsaufwand minimiert werden und die zeitliche sowie finanzielle Herausforderung für Lehrer*innen und Schulen, sechs Spielexemplare eigenständig zu produzieren und nach jedem Einsatz erneut vorzubereiten, vollständig entfallen. Breuer hat festgestellt, dass Lehrer*innen bei der Implementierung neuer didaktischer Instrumente in ihren Unterricht zögerlich reagieren und es daher notwendig ist, ihnen die Neuerungen möglichst niederschwellig anzubieten, um eine erste Implementierung und einen nachfolgenden breiten Einsatz im Unterricht überhaupt erst zu ermöglichen [10]. Weiterhin könnte das Spiel durch den Einsatz während der am Institut stattfindenden Strahlenschutzkurse für Lehrer*innen Überzeugungsarbeit leisten. Wie Thurm festgestellt hat, beeinflusst die Selbstwahrnehmung der Lehrer*innen in Fortbildungsmaßnahmen maßgeblich den späteren Einsatz der erlernten Fortbildungsinhalte im Unterricht [62]. Die Notwendigkeit der größeren Anzahl an Spielexemplaren und die damit verbundenen Vorschläge zur weiterführenden Entwicklung eines Implementierungskonzeptes würden den Lehrer*innen und Schüler*innen zu Gute kommen und stellen aus Sicht des Versuchsleiters die einzig wirklich praktikable Möglichkeit dar, das Spiel erfolgreich in den Physikunterricht einzubinden.

Diese Arbeit stellt zweifelsfrei nicht den Abschluss der Untersuchungsmöglichkeiten des verwendeten EXIT-Spiels dar. Anschließende Studien beispielsweise zum Einfluss der Lehrkraft und der Passung des Unterrichts zu den Spielinhalten können genauso Informationen liefern, die eine angemessenere Beurteilung des Spieleinsatzes im Unterricht erlauben, wie Untersuchungen zu geförderten prozessbezogenen Kompetenzen oder der Rolle des Versuchsleiters während der Datenerhebung. Besonders letzterer Analyseaspekt stellt sich als interessant heraus, da der Versuchsleiter dieser Erhebung gleichzeitig der Entwickler des Spiels war. So könnten potentielle Effekte durch die besondere Fachkunde im Bereich des Spiels die Daten durch bewusste und unbewusste Interventionen im Vergleich zu einer regulären Lehrkraft, die das Spiel einsetzt, verzerrt haben. All diese Aspekte könnten durch weiterführende Erprobungen und stetige Verbesserungen des Spiels im Schuleinsatz untersucht werden.

Letztendlich kann an dieser Stelle nur festgehalten werden, dass wenn auch kein signifikanter Lerneffekt des Spiels nachgewiesen werden konnte, zumindest die Motivation der Schüler*innen für die Thematisierung mit physikalischen Rätseln und Problemstellungen nichtsdestotrotz hoch ausfiel. Auch wenn die Schüler*innen ihr Fachwissen zur Kernphysik nicht merklich steigern konnten, so wurde zumindest bestehendes Wissen durch den Spieleinsatz wiederholt. Abschließend kann also trotz und auch gerade wegen der Untersuchungsergebnisse nur für den Einsatz von Spielen und explizit diesem EXIT-Spiel im Physikunterricht appelliert werden, wie es auch das nachfolgende Zitat von Mikelskis-Seifert deutlich belegt. *„Gerade die Freude und der Spaß können eine positive Einstellung der Schüler zum Physikunterricht bewirken und die emotionale Distanz zum Physiklehrer abbauen […].“* [14].

10. Danksagung

Eine Vielzahl von Personen hat diese Arbeit durch ihr Mitwirken überhaupt erst möglich gemacht, weshalb ihnen an dieser Stelle gedankt werden soll.

Als erstes möchte ich mich hier selbstverständlich bei meinem Erstprüfer Prof. Dr. Clemens Walther, dem Leiter des Instituts für Radioökologie und Strahlenschutz der Leibniz Universität Hannover bedanken. Ohne seine tatkräftige Unterstützung, sowie seine dauerhafte und spontane Erreichbarkeit, ob im Urlaub, Krankheit oder für eine E-Mail zu später Stunde, wäre diese Arbeit nicht denkbar gewesen.

Gleichwohl sei auch meinem Zweitgutachter Dr. Jan-Willem Vahlbruch, vor allem für die unproblematische und zügige Organisation der Auswertungssoftware, sowie der Unterstützung bei der Entwicklung des Erhebungsinstruments, gedankt.

An dritter Stelle gebührt mein großer Dank Prof. Dr. Gunnar Friege, dem Leiter der AG Physikdidaktik an der Leibniz Universität Hannover. Auch wenn er selbst und die AG Physikdidaktik nichts mit dieser Abschlussarbeit zu tun hatten, stand er mir dennoch bei jeder Frage und allen Problemen, die im Laufe des letzten halben Jahres aufgetreten sind, zur Seite und hatte immer einen passenden Ratschlag zur Hand.

Neben der Betreuung meiner Arbeit am Institut für Radioökologie und Strahlenschutz habe ich einen Großteil der Bearbeitungszeit zur Datenerhebung an zwei Schulen in der Hannoveraner Südstadt verbracht. Aus diesem Grund gebührt mein Dank den beiden Physiklehrern Robert Huckemann von der St. Ursula-Schule und Markus Wießell von der Bismarckschule, die von Beginn an von der Forschungsidee und dem Spiel begeistert gewesen sind und mir ohne zu zögern über einen Monat ihrer Unterrichtszeit in insgesamt vier Klassen eingeräumt haben. Beide haben keine Mühe gescheut auch spontan auf unvorhergesehene Schwierigkeiten vor und während der Datenerhebung zu reagieren und mir so an der ein oder anderen Stelle aus der Misere geholfen.

Zu den beiden Physiklehrern gehören natürlich auch vier Klassen des 10. Jahrganges, die aus Datenschutzgründen nur anonymisiert erwähnt werden. Vielen Dank für Eure Ausdauer über einen Monat hinweg einem für euch vollkommen Fremden dabei zu helfen seine Masterarbeit zu schreiben.

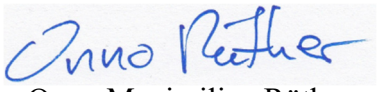
Auch den Schulleiter*innen beider Schulen danke ich für die unkomplizierte und zügige Genehmigung der Datenerhebung.

Die langen Tage der oftmals sehr ermüdenden statistischen Datenauswertung im Institut wurden durch jedes kleine Gespräch mit den Kolleg*innen ein wenig erträglicher. Ich

danke Euch allen für jede morgentliche Kaffeerunde, jeden Ausflug in der Mittagspause zur Mensa und noch viel mehr, dass ihr immer hilfsbereit gewesen seid und mir bei Problemen egal welcher Art mit Rat und Tat zur Seite standet.
Gleichwohl sei natürlich meiner Familie und all meinen Freunden gedankt, dass Ihr es ein halbes Jahr mit meinem Masterarbeits-Ich ausgehalten habt, dass exponentiell anstrengender als meine ohnehin schon herausfordernde Persönlichkeit gewesen sein muss. Vielen Dank für alle aufmunternden Worte, jede Art der Ablenkung um einfach mal abschalten zu können und natürlich jeder Hilfe und am Ende auch Korrektur der Ergebnisse meiner Arbeit. Eine Person aus diesem Kreise muss zum Abschluss allerdings noch namentlich erwähnt werden, da ich ohne ihn sicherlich nach der Hälfte der Bearbeitungszeit das Handtuch geworfen hätte.
Mein größter Dank gebührt dir lieber Simon! Ohne deine Erfahrungen und dein Knowhow im Umgang mit SPSS oder R-Studio, ohne die telefonischen Ferndiagnosen und Lösungsversuche meiner Probleme in der statistischen Auswertung der Daten, ohne die stundenlange Korrektur meiner Bandwurmsätze, ohne ein einziges Komma, ohne die ein oder andere abendliche Radtour mit anschließendem Feierabendbier und deinem unermüdlichen Talent mir stundenlang zuzuhören und nichtsdestotrotz immer noch mit mir befreundet zu sein, ohne all dies wäre diese Arbeit nie möglich gewesen. Vielen Dank für Alles!

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt worden ist.



Hannover, den 17.08.2022 Onno Maximilian Rüther

Literaturverzeichnis

[1] BRAUN, Cornelia: *Spielend lernen – Digitale Spiele in der schulischen Medienbildung.* (Masterarbeit Kultur- und Medienwissenschaften), Hochschule Merseburg, Leipzig 2016.

[2] IMHOF, Margarete: *Psychologie für Lehramtsstudierende.* Springer Verlag, Wiesbaden ^{4}2016.

[3] HÄRTIG, Hendrik: *Im Physikunterricht Spielen! Charakteristika von Spielen und Chancen für den Physikunterricht.* in: Unterricht Physik, Jg. 2015, Nr. 149, 2-5.

[4] KIPMAN, Ulrike: *Problemlösen Begriff-Strategien-Einflussgrößen-Unterricht-(häusliche) Förderung*. Springer Gabler Verlag. Wiesbaden ^{2}2020.

[5] HEINICKE, Susanne / HOLZ, Christoph: *Wann wird man aus Fehlern klug? Perspektiven auf den Umgang mit und das Lernen aus Fehlern.* in: Unterricht Physik, Jg. 2020, Nr. 177/178, 4-9.

[6] HEPP, Ralph / MIEHE, Kirsten: *Kooperatives Lernen trainieren. Hinweise und Empfehlungen für den Einstieg in kooperative Lernformen.* in: Unterricht Physik, Jg. 2004, Nr. 84, 8-13.

[7] HEPP, Ralph: *Wie kann üben gelingen? Methoden und Strategien nachhaltigen Übens.* in: Unterricht Physik, Jg. 2019, Nr. 173, 8-11.

[8] HEPP, Ralph: *Kein Lernen ohne Üben. Effektives Lernen initiieren und Gelerntes nachhaltig sichern.* in: Unterricht Physik, Jg. 2019, Nr. 173, 2-7.

[9] RÜTHER, Onno Maximilian: *Entwicklung eines didaktischen Instruments zur Festigung von Lehrinhalten zu Radioaktivität in der Sekundarstufe II.* (Bachelorarbeit Physik), Gottfried Wilhelm Leibniz Universität, Hannover 2020.

[10] BREUER, Judith / VOGELSANG, Christoph / REINHOLD, Peter: *Nutzungsverhalten von Lehrkräften bei der Implementierung einer physikdidaktisch innovativen Unterrichtskonzeption.* in: Zeitschrift für Didaktik der Naturwissenschaften, Jg. 2022, Nr. 28, 1-13. [online]: https://doi.org/10.1007/s40573-022-00138-5 [07.07.2022].

[11] GRANDE-DE-PRADO, Mario / GARCÍA-MARTÍN, Sheila / BAELO, Roberta / ABELLA-GARCÍA, Víctor: *Edu-Escape Rooms.* in: Encyclopedia 2021, 1, 12-19. [online]: https://doi.org/10.3390/encyclopedia1010004 [16.05.2022].

[12] RICKEN, Gabi: *Lernprozessdiagnostik.* in: ARNOLD, Karl-Heinz / SANDFUCHS, Uwe / WIECHMANN, Jürgen (Hrsgg.): *Handbuch Unterricht.* Verlag Julius Klinkhardt, Bad Heilbrunn ^{2}2009, 476-479.

[13] GRUBBAUER, Michaela: *Spielen als pädagogische Maßnahme. Präventive, spielorientierte Förderung und Stärkung elterlicher Kompetenz*. VS Verlag für Sozialwissenschaften, Wiesbaden 2011.

[14] MIKELSKIS-SEIFERT, Silke / BEHRENDT, Helga: *Spielen im Physikunterricht.* in: MIKELSKIS-SEIFERT, Silke / RABE, Thorid (Hrsgg.): *Physik Methodik. Handbuch für die Sekundarstufe I und II*. Cornelsen Scriptor Verlag, Berlin 2010.

[15] MEYER, Hilbert: *Unterrichtsmethoden II – Praxisband*. Cornelsen Verlag, Berlin ^{16}2020.

[16] FÜRSTENAU, Bärbel: *Planspiel und Simulation.* in: ARNOLD, Karl-Heinz / SANDFUCHS, Uwe / WIECHMANN, Jürgen (Hrsgg.): *Handbuch Unterricht.* Verlag Julius Klinkhardt, Bad Heilbrunn ^{2}2009, 240-243.

[17] AUER, Verena: *Spielen im Physikunterricht.* in: Delta Phi B. 2015. [online]: http://www.physikdidaktik.info/index.php/Delta_Phi_B_2015 [17.05.2022].

[18] VELDKAMP, Alice / VAN DE GRINT, Liesbeth / KNIPPELS, Marie-Christine P. J. / VAN JOOLINGEN, Wouter R.: *Escape Education: A systematic review on escape rooms in education.* in: Educational Research Review, Volume 31, 2020. [online]: https://doi.org/10.1016/j.edurev.2020.100364 [16.05.2020].

[19] VELDKAMP, Alice / KNIPPELS, Marie-Christine P. J. / VAN JOOLINGEN, Wouter R.: *Beyond the Early Adopters: Escape Rooms in Science Education.* in: frontiers of Education, 2021. [online]: https://doi.org/10.3389/feduc.2021.622860 [20.06.2022].

[20] PRAX, Lukas: *Motivation im Unterricht.* in: Delta Phi B. 2016. [online]: http://www.physikdidaktik.info/index.php/Delta_Phi_B_2016 [07.07.2022].

[21] VELDKAMP, Alice / DAEMEN, Joke / TEEKENS, Stijn / LOELEWIJN, Stefan / KNIPPELS, Marie-Christine P. J. / VAN JOOLINGEN, Wouter R.: *Escape boxes: Bringing escape room experience into the classroom.* in: British Journal of Educational Technology, Volume 51, Number 4, 2020, 1220-1239. [online]: https://doi.org/10.1111/bjet.12935 [21.06.2022].

[22] WIEMKER, Markus / ELUMIR, Errol / CLARE, Adam: *Escape Room Games.* in: HAAG, Johann / WEISSENBÖCK, Josef / GRUBER, Wolfgang / FREISLEBEN-TEITSCHER, Christian F. (Hrsgg.): *Game Based Learning. Dialogorientierung & spielerisches Lernen analog und digital.* Beiträge zum 4. Tag der Lehre an der FH St. Pölten am 15.10.15, St. Pölten 2015.

[23] SANCHEZ, Eric / PLUMETTAZ-SIEBER, Maud: *Teaching and Learning with Escape Games from Debriefing to Institutionalization of Knowledge.* in: GENTILE, Manuel / ALLEGRA, Mario / SÖBKE, Heinrich (Hrsgg.): *Games and Learning Alliance.* 7th International Conference, GALA 2018, Springer Nature, Cham 2019, 242-253. [online]: https://doi.org/10.1007/978-3-030-11548-7_23 [22.06.2022].

[24] FOTARIS, Panagiotis / MASTORAS, Theodoros: *Escape Rooms for Learning: A Systematic Review.* Proceedings of the 13th International Conference on Game Based Learning, ECGBL 2019, 235-243.

[25] O`BRIEN, Kelsey / PITERA, Jenna: *Gamifying Instruction and Engaging Students With Breakout EDU.* in: Journal of Educational Technology Systems, Volume 48, Issue 2, 2019, 192-212. [online]: https://doi.org/10.1177/0047239519877165 [22.06.2022].

[26] MAKRI, Agoritsa / VLACHOPOULOS, Dimitrios / MARTINA, Richard A.: *Digital Escape Rooms as Innovative Pedagogical Tools in Education: A Systematic Literature Review.* in: Sustainability 2021, 13, 4587. [online]: https://doi.org/10.3390/su13084587 [20.06.2022].

[27] SÁRKÖZI, Zsuzsa / BORBÉLY, Sándor / JÁRAI-SZABÓ, Ferenc: *Deepening Secondary Students Understanding of Physics through Escape Games.* AIP Conference Proceedings 2071. 2019. [online]: https://doi.org/10.1063/1.5090085 [20.06.2022].

[28] HOPF, Martin / SCHECKER, Horst: *Unterrichtskonzeptionen zu fortgeschrittenen Themen der Schulphysik.* in: WILHELM, Thomas / SCHECKER, Horst / HOPF, Martin (Hrsgg.): *Unterrichtskonzeptionen für den Physikunterricht. Ein Lehrbuch für Studium, Referendariat und Unterrichtspraxis.* Springer Spektrum Verlag, Berlin 2021, 369-400.

[29] MIJAL, Michał / CIEŚLA, Martyna / GROMADZKA, Monika: *Educational Escape Room. Challenges and Obstacles.* in: WARDASZKO, Marcin / MEIJER, Sebastiaan / LUKOSCH, Heide / KANEGAE, Hidehiko / KRIZ, Willy Christian

/ GRZYBOWSKA-BRZEZIŃSKA, Mariola (Hrsgg.): *Simulation Gaming Through Times and Disciplines.* 50th International Simulation and Gaming Association Conference, ISAGA 2019, Springer Nature, Cham 2021, 84-98. [online]: https://doi.org/10.1007/978-3-030-72132-9_8 [21.06.2022].

[30] VITA VÖROS, Alpár István / SÁRKÖZI, Zsuzsa: *Physics Escape Room as an Educational Tool.* AIP Conference Proceedings 1916, 2017. [online]: https://doi.org/10.1063/1.5017455 [20.06.2022].

[31] NIEDERSÄCHSICHES KULTUSMINISTERIUM (Hrsg.): *Kerncurriculum für das Gymnasium - gymnasiale Oberstufe, die Gesamtschule - gymnasiale Oberstufe, das Berufliche Gymnasium, das Abendgymnasium, das Kolleg. Physik*, Hannover 2022. [online]: https://cuvo.nibis.de/cuvo.php?p=search& [21.07.2022].

[32] GRASSINGER, Robert / DICKHÄUSER, Oliver / DRESEL, Markus: *Motivation*. in: URHAHNE, Detlef / DRESEL, Markus / FISCHER, Frank (Hrsgg.): *Psychologie für den Lehrberuf.* Springer Verlag, Berlin 2019, 207-227.

[33] WILDE, Matthias / BÄTZ, Katrin / KOVALEVA, Anastassiya / URHAHNE, Detlev: *Überprüfung einer Kurzskala intrinsischer Motivation (KIM).* in: Zeitschrift für Didaktik der Naturwissenschaften, Jg. 2009, Nr. 15, 31-45.

[34] HAMARI, Juho / SHERNOFF, David J. / ROWE, Elizabeth / COLLER, Brianno / ASBELL-CLARKE, Jodi / EDWARDS, Teon: *Challenging games help students learn: An empirical study on engagement, flow an immersion in game-based-learning.* in: Computers in Human Behavior, Volume 54, 2016, 170-179. [online]: https://doi.org/10.1016/j.chb.2015.07.045 [20.06.2022].

[35] WELLENREUTHER, Martin: *Forschungsbasierte Schulpädagogik. Anleitung zur Nutzung empirischer Forschung für die Schulpraxis.* Schneider Verlag Hohengehren, Baltmannsweiler 42014.

[36] LAZONDER, Ard W. / HARMSEN, Ruth: *Meta-Analysis of Inquiry-Based Learning: Effects of Guidance.* in: Review of Educational Research, Volume 86, Issue 3, 681-718. [online]: https://doi.org/10.3102/0034654315627366 [20.06.2022].

[37] SKENE, Kayleigh / O'FARRELLY, Christine M. / BYRNE, Elizabeth M. / KIRBY, Natalie / STEVENS, Eloise C. / RAMCHANDANI, Paul G.: *Can guidance during play enhance children's learning and development in educational contexts? A systematic review and meta-analysis.* in: Child Development, 00, 1–19. [online]: https://doi.org/10.1111/cdev.1373 [16.05.2022].

[38] DÖRING, Nicola / BORTZ, Jürgen: *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften.* Springer Verlag, Berlin ^{5}2016.

[39] BIERHOFF, Hans Werner / RUDINGER, Georg: *Quasiexperimentelle Untersuchungsmethoden*. in: ERDFELDER, Edgar / MAUSFELD, Rainer / MEISER, Thorsten / RUDINGER, Georg (Hrsgg.): *Handbuch quantitative Methoden*. Beltz Psychologie Verlags Union, Weinheim 1996, 47-58.

[40] LEISEN, Josef / HÖTTECKE, Dietmar: *Leistungsmessung und Schülerbeurteilung.* in: WIESNER, Hartmut / SCHECKER, Horst / HOPF, Martin (Hrsgg.): *Physikdidaktik kompakt.* Aulis-Verlag, Hallbergmoos 2011, 63-71.

[41] NIEDERSÄCHSICHES KULTUSMINISTERIUM (Hrsg.): *Kerncurriculum für das Gymnasium Schuljahrgänge 5-10. Physik*, Hannover 2015. [online]: https://cuvo.nibis.de/cuvo.php?p=search& [21.07.2022].

[42] NIEDERSÄCHSICHES KULTUSMINISTERIUM (Hrsg.): *17. Physik - Hinweise zur schriftlichen Abiturprüfung 2022*, Hannover 2021. [online]: https://www.nibis.de/2022_11938 [21.07.2022].

[43] HÖTTECKE, Dietmar / WODZINSKI, Rita: *Diagnostizieren und Fördern. Hintergründe, Ansätze und Probleme von Diagnostik im Physikunterricht.* in: Unterricht Physik, Jg. 2015, Nr. 147/148, 2-10.

[44] HÄUSSLER, Peter: *Wie lässt sich der Lernerfolg messen?* in: KIRCHER, Ernst / GIRWIDZ, Raimund / HÄUSSLER, Peter (Hrsgg): *Physikdidaktik. Theorie und Praxis*. Springer Spektrum Verlag, Berlin/Heidelberg ^{3}2015, 247-293.

[45] WORBACH, Marc / DRECHSEL, Barbara / CARSTENSEN, Claus H.: *Messen und Bewerten von Lernergebnissen*. in: URHAHNE, Detlef / DRESEL, Markus / FISCHER, Frank (Hrsgg.): *Psychologie für den Lehrberuf.* Springer Verlag, Berlin 2019, 493-516.

[46] SACHER, Werner: *Lernstandsbeurteilung: Tests, Zensuren, Zeugnisse.* in: ARNOLD, Karl-Heinz / SANDFUCHS, Uwe / WIECHMANN, Jürgen (Hrsgg.): *Handbuch Unterricht.* Verlag Julius Klinkhardt, Bad Heilbrunn ^{2}2009, 483-490.

[47] MAGILL, Joseph / DREHER, Raymond / SÓTI, Zslot: *Karlsruher Nuklidkarte*. Nucleonica GmbH, Karlsruhe ^{10}2018.

[48] PARADIES, Liane / WESTER, Franz / GREVING, Johannes (Hrsgg.): *Leistungsmessung und -bewertung.* Cornelsen Scriptor Verlag, Berlin ^{5}2014.

[49] BÜHNER, Markus: *Einführung in die Test- und Fragebogenkonstruktion.* Pearson Studium Verlag, München ^{3}2011.

[50] HOPF, Martin / SCHECKER, Horst: *Schülervorstellungen zu fortgeschrittenen Themen der Schulphysik.* in: SCHECKER, Horst / WILHELM, Thomas / HOPF, Martin / DUIT, Reinders (Hrsgg.): *Schülervorstellungen und Physikunterricht, Ein Lehrbuch für Studium, Referendariat und Unterrichtspraxis.* Springer Spektrum Verlag, Berlin 2018, 225-242.

[51] FISCHER, Hans E. / KRABBE, Heiko: *Empirische Forschung in der Physikdidaktik.* in: KIRCHER, Ernst / GIRWIDZ, Raimund / HÄUSSLER, Peter (Hrsgg): *Physikdidaktik. Theorie und Praxis*. Springer Spektrum Verlag, Berlin/Heidelberg ^{3}2015, 727-757.

[52] SACHER, Werner: *Leistungen entwickeln, überprüfen und beurteilen. Bewährte und neue Wege für die Primär- und Sekundarstufe.* Verlag Julius Klinkhardt, Bad Heilbrunn ^{5}2009.

[53] FIELD, Andy: *Discovering Statistics using SPSS.* Sage Publications, London ^{3}2009.

[54] JANSSEN, Jürgen / LAATZ, Wilfried: *Statistische Datenanalyse mit SPSS. Eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests.* Springer Gabler Verlag, Berlin ^{9}2017.

[55] KRONTHALER, Franz: *Statistik angewandt. Datenanalyse ist (k)eine Kunst. Excel Edition.* Springer Spektrum Verlag, Berlin/Heidelberg 2016.

[56] SEDLMEIER, Peter / RENKEWITZ, Frank: *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler*, Pearson Verlag, Hallbergmoos ^{2}2013.

[57] SCHWAIGHOFER, Matthias / HEENE, Moritz / BÜHNER, Markus: *Grundlagen und Kriterien der Diagnostik.* in: URHAHNE, Detlef / DRESEL, Markus / FISCHER, Frank (Hrsgg.): *Psychologie für den Lehrberuf.* Springer Verlag, Berlin 2019, 471-491.

[58] HÖTTECKE, Dietmar: *Stolpersteine der Diagnostik … und wie man sie umgehen kann.* in: Unterricht Physik, Jg. 2015, Nr. 147/148, 11-13.

[59] BUCHNER, Alex / ERDFELDER, Edgar / FAUL, Franz: *Teststärkeanalysen*. in: ERDFELDER, Edgar / MAUSFELD, Rainer / MEISER, Thorsten / RUDINGER, Georg (Hrsgg.): *Handbuch quantitative Methoden*. Beltz Psychologie Verlags Union, Weinheim 1996, 123-136.

[60] WODZINSKI, Rita: *Kooperatives Lernen: mehr als nur Gruppenarbeit Gründe für kooperatives Arbeiten im Physikunterricht.* in: Unterricht Physik, Jg. 2004, Nr. 84, 4-7.

[61] RAULIN, Dennis: *Radlab, Ein mobiles Schülerlabor zum Thema Radioaktivität.* (Masterarbeit Physik), Gottfried Wilhelm Leibniz Universität, Hannover 2022.

[62] THURM, Daniel: *Digitale Werkzeuge im Mathematikunterricht integrieren. Zur Rolle von Lehrerüberzeugungen und der Wirksamkeit von Fortbildungen.* Springer Spektrum Verlag, Wiesbaden 2020.

Anhang

Der Anhang dieser Arbeit ist der beigefügten CD zu entnehmen. Dieser sind neben allen Materialien des Spiels nicht nur die bearbeiteten Datensätze, sondern auch die Rohdaten der Erhebungen zu entnehmen.