Entwicklung und Validierung des Luxemburger Orthografietests

Abschlussbericht Dezember 2023

Dr. Philipp Sonnleitner
Ulrich Keller
Helmuth Sperl

Aufbau und Inhalt

Einleitung und Anforderungsprofil

Die Entwicklung der letzten Jahre zeigt eine stetig wachsende Anzahl an Luxemburgisch-Lernenden* und die zunehmende Verwendung des Luxemburgischen im privaten und administrativen Schriftverkehr. Trotz der damit großen Relevanz der luxemburgischen Schriftsprache, steht bis dato aber kein standardisiertes und validiertes Instrument zur Erfassung der Orthografiekenntnisse zur Verfügung. Ein derartiges Verfahren wäre aber in vielerlei Hinsicht von Vorteil wenn nicht sogar notwendig, sei es zur Evaluierung von Weiterbildungen und Kursen, oder zur Prüfung von Qualifikation für gewisse Positionen. Bisher wurden Rechtschreibkenntnisse beispielsweise an der Universität Luxemburg oder im Rahmen von Luxemburgischkursen in der Erwachsenenbildung mithilfe von Diktaten evaluiert.

Die Erstellung der Diktate, als auch deren Korrektur bedeutet allerdings einen erheblichen (Zeit-)Aufwand. Angesichts der prognostizierten Zunahme an notwendigen Testungen der Orthografiekenntnisse, stößt das Diktatformat vor allem auch an eine Durchführbarkeitsgrenze was die Testadministration betrifft. Je mehr getestet wird, desto mehr Testversionen sind notwendig, um das Bekanntwerden der Testinhalte erfolgreich zu verhindern. Auch ist davon auszugehen, dass die Testungen zunehmend dezentral, also nicht nur an den etablierten Schulungsinstituten (INLL, ZLS, Universität Luxemburg) durchgeführt werden, sondern beispielsweise auch in Schulklassen oder kleineren Gemeinden, was die Standardisierung der Testdurchführung weiters gefährdet.

Diese Ausgangsbasis diente dazu, ein Anforderungsprofil für einen neuen, standardisierten und psychometrisch validierten Orthografietest zu erstellen. In Zusammenarbeit zwischen ZLS und dem LUCET wurden folgende Aspekte festgelegt:

Testentwicklung (Juli 2021 – Februar 2022)	1.1	Leichte und benutzerfreundliche Durchführbarkeit am Computer/ Tablet
	1.2	Möglichkeit zur Durchführung innerhalb einer Schulstunde
	1.3	Geschlossenes Antwortformat zur Gewährleistung objektiver und standardisierter, damit automatisierbarer Korrektur
	1.4	Erstellung ausreichend diverser Testversionen um Testsicherheit zu gewährleisten
	1.5	Inhaltliche Validität durch Abdeckung der wichtigsten Orthografieregeln
Pilotierung (März 2022 – September 2022)	2.1	Eindimensionalität, also psychometrisch abgesicherte Verwendung eines Summenscores als reliablen Indikator der Orthografiekenntnisse
	2.2	Fairness hinsichtlich demografischer Merkmale wie Geschlecht, Alter, regionaler Herkunft
	2.3	Gute Akzeptanz unter der Zielgruppe des Tests
Validierung (Oktober 2022 – August 2023)	3.1	Konstruktvalidität durch Übereinstimmung mit bisher eingesetzten Verfahren

*Es wurde versucht, stets geschlechtsneutral zu formulieren, ansonsten sind beim generischen Maskulinum Frauen natürlich mitgemeint

Anhand dieses Anforderungsprofils wurde ein Fahrplan entwickelt, der die Testentwicklung und Implementierung in der am LUCET entwickelten online Testplattform OASYS vorsah (1.1 bis 1.5). Darauf aufbauend wurde eine Pilotstudie zur Validierung und Kalibrierung der entwickelten Testaufgaben (= Items) durchgeführt (2.1 bis 2.3) und abschließend wurde die finale Testform in einer Validierungsstudie überprüft (3.1). Der vorliegende Bericht orientiert sich an dieser Einteilung und dokumentiert sogleich Entwicklung und Validierung des luxemburgischen Orthografietests.

Testentwicklung

Testframework

In enger Abstimmung zwischen ZLS und LUCET wurde zuerst ein provisorisches Testframework vereinbart, das die zugehörige Aufgabenentwicklung strukturieren sollte. Als Ausgangsbasis diente dabei das Standardwerk D'Lëtzebuerger Orthografie (CPLL & ZLS, 2021). In einem ersten Schritt wurden die getesteten Hauptkategorien (z.B. *1. D'Vokaler a, i, o an u*) und Subkategorien festgelegt (z.B. 1.1. *D'Quantitéitsreegel*). Danach wurde die Relevanz der einzelnen Kategorien für den Test bestimmt und für eine erste angenommene Testlänge mit 110 Aufgaben aufgeschlüsselt. Diese Gewichtung sollte einerseits die Wichtigkeit der jeweiligen Orthografieregel repräsentieren, andererseits der Häufigkeit der betroffenen Wörter entsprechen. Tabelle 1 zeigt die dabei festgelegte Auswahl an Hauptkategorien mit der jeweiligen Gewichtung für die Entwicklung. Die jeweiligen Subkategorien sollten möglichst ausgewogen repräsentiert sein.

Obwohl nicht als eigene (Haupt-)Kategorie ausgewiesen, wurde bei der Itementwickung besonderes Augenmerk auf so genannte „Exoten" gelegt, Wörter, die zwar in einem relevanten Ausmaß in der Schriftsprache vorkommen, teils aber von allgemeinen Regeln abweichen oder sehr spezifisch sind und deshalb als besonders schwierig gelten. Diese sollten zumindest mit 5 Items im Test, über die anderen Kategorien verteilt, repräsentiert sein.

Tabelle 1: Gewichtung der einzelnen Hauptkategorien für die Itementwicklung

Hauptkategorie	**Gewichtungsfaktor**	**Anzahl Items in Test**
D'Vokaler a, i, o an u	3	15
De Vokal e	3	15
D'Konsonanten	3	15
D'Verben	2	10
D'Friemwierder	2	10
D'n-Reegel	2	10
D'r-Reegel	2	5
Vokalkoppelen	2	5
D'Grouss- a Klengschreiwung	1	5
D'Getrennt- an Zesummeschreiwung	1	5
Exoten	1	5
Total	**22**	**110**

Itementwicklung

Auf Basis des Testframeworks (siehe Tabelle 1) wurde am ZLS mit der Itementwicklung begonnen. Ein durch das LUCET durchgeführter Workshop am ZLS (17.09.2021), führte allgemeine Prinzipien und Empfehlungen zur Itementwicklung ein (siehe z.B. Downing & Haladyna, 2006), die in einem zweiten Workshop am 15.10.2021 vertieft und im weiteren Ablauf berücksichtigt wurden. Ziel war es, das Testframework möglichst ausgewogen in Bezug auf die abgetesteten Kompetenzen abzudecken. Für jede entwickelte Aufgabe, wurde vonseiten der Itemersteller ein standardisiertes Excel-Sheet ausgefüllt, das die folgenden Punkte umfasste:

- Durch das Item abgetestete Hauptkategorie
- Durch das Item abgetestete Subkategorie
- Beispiel für die abgetestete Rechtschreibregel
- Itemstamm (Satz mit Wortlücke)
- Korrekte Antwort
- 3 Distraktoren (plausible aber eindeutig falsche Alternativantworten)
- Eingeschätzte Schwierigkeit des Items
- Kommentar bzw. Frage an die Reviewer

Jedes entwickelte Item wurde wiederum durch zwei unabhängige Reviewer am ZLS kritisch gegengelesen und geprüft. Dafür diente eine Art Leitfaden, der folgende Punkte umfasste:

- Eindeutigkeit der Zuordnung zu den Kompetenzkategorien
- Richtigkeit der Lösung und Inkorrektheit der Distraktoren
- Überschneidung von Items
- Gefahr von unterschiedlichem Funktionieren der Aufgabe je nach demografischem Hintergrund (Differential item functioning, DIF). Dies könnte vor allem durch Unterschiede in der Aussprache/Verwendung mancher Wörter je nach Alter, Herkunft oder sprachlichem Hintergrund zutreffen.
- Orthografie des Itemstamms
- Einfachheit des formulierten Satzes
- Sinnhaftigkeit des abgeprüften Wortes (Frequenz im Sprachgebrauch) und zugehöriger Kategorie

Die Entwicklung und das anschließende Review folgten einem sehr zeitintensiven iterativen Prozess, der Änderungen oder das Löschen einzelner Items inkl. erneutem Review nach sich zog. Waren die Einschätzungen der beiden Reviewer unterschiedlich, gab es eine inhaltliche Diskussion teils unter Beiziehung eines dritten Reviewers. Ebenfalls gab es durch diesen Prozess für die Entwicklung spezifische Definitionsänderungen der abgetesteten Kompetenzen. Am 21.02.2022 stand der finale Korpus mit insgesamt 1037 Items. Anzumerken ist hierbei, dass deutlich mehr Items entwickelt wurden, diese jedoch aus unterschiedlichen Gründen beim Review ausgeschieden wurden. Die Größe des Itempools kann jedenfalls als bemerkenswert angesehen werden und bildete eine ideale Ausgangslage für die Pilotierung des Orthografietests.

In den nachfolgenden Tabellen 2 und 3 ist die Verteilung der Aufgaben auf die Kompetenzbereiche, als auch die Einschätzung der Reviewer hinsichtlich Schwierigkeit und Risiko für DIF angegeben. Diese Informationen lieferten eine gute Grundlage für den nächsten Schritt: das Design der Pilotierung.

Tabelle 2: Anzahl entwickelter Items pro Hauptkategorie

Hauptkategorie	**Häufigkeit**	**Prozent**
D'Vokaler a, i, o an u	106	10,2
De Vokal e	81	7,8
D'Konsonanten	221	21,3
D'Verben	79	7,6
D'Friemwierder	165	15,9
D'n-Reegel	148	14,3
D'r-Reegel	65	6,7
Vokalkoppelen	46	4,4
D'Grouss- a Klengschreiwung	79	7,6
D'Getrennt- an Zesummeschreiwung	47	4,5
Exoten	n.a.	n.a.
Total	**1037**	**100**

*Im Orthografietest wurden Items der Kategorie „Exoten" auf die anderen Kategorien aufgeteilt

Tabelle 3: Einschätzung der Items hinsichtlich Risiko für Differential Item Functioning (DIF) und Schwierigkeit

Attribut		**Häufigkeit**	**Prozent**
DIF-Risiko	ja	30	2,9
	nein	1007	97,1
Schwierigkeit	einfach	596	57,5
	mittel	344	33,2
	schwierig	97	9,4

Pilotierungsstudie

Zielsetzung

Die sehr hohe Anzahl an entwickelten Testaufgaben hatte mehrfache Implikationen für die Pilotierungsstudie, deren Ziel es war, möglichst alle entwickelten Items anhand möglichst vieler und diverser Teilnehmer zu testen. Einerseits konnte man für die Durchführung der Pilotierung aus dem Vollen schöpfen und die Abdeckung der Orthografiekompetenz gut inhaltlich abdecken. Andererseits stellte diese Fülle an Items hohe Anforderungen an das Testdesign, damit auch sichergestellt war, dass jede Aufgabe von genügend Teilnehmern bearbeitet wird, um reliable und zuverlässige Schätzungen der Itemparameter zu bekommen. Am Ende sollten zusätzlich zur psychometrischen Kalibrierung des entwickelten Itempools Erfahrungswerte für die Gestaltung des Orthografietests hinsichtlich Länge, Dauer, Schwierigkeit und Fairness gewonnen werden.

Methode

Testdesign und Simulationsstudie

Das ursprünglich angedachte unvollständige Blockdesign, bei dem mehrere, sich teils überlappende („verlinkte“) Testversionen administriert werden, stieß sehr schnell an seine Grenzen. Es war rasch absehbar, dass mit diesem Ansatz nicht alle entwickelten Items gepretestet werden konnten, oder aber die benötigte Stichprobe sehr groß werden würde. Während die erste Option angesichts der sehr erfolgreichen Itementwicklung nur suboptimal gewesen wäre, so würde die zweite Option ganz einfach sehr rasch an Machbarkeitsgrenzen stoßen, da die Rekrutierung von motivierten und geeigneten Teilnehmern ohnehin schon eine Herausforderung darstellte. Das LUCET versuchte deshalb einen neuen Ansatz zur Erstellung von verschiedenen Testversionen.

Zuerst wurden aufgrund der Verteilung der Items auf die Hauptkategorien, randomisiert kompetenzübergreifende 5er-Blöcke erstellt. Obwohl die empirische Verteilung nicht exakt die vorgesehene Gewichtung des Testframeworks wiedergibt, stand bei der Pilotierung vor allem eine möglichst umfassende Erprobung der entwickelten Aufgaben im Vordergrund. Durch diesen Ansatz entstanden 207 vorgefertigte Testseiten à 5 Items. Dabei wurde darauf geachtet, dass die Schwierigkeit der einzelnen Blöcke vergleichbar ist: die eingeschätzte Schwierigkeit beider Reviewer wurde gemittelt und in Relation zu den Kompetenzen gesetzt. Dabei entstanden bis auf einzelne Blöcke ausschließlich Testseiten mit mittlerer Schwierigkeit, weswegen bei der nachfolgenden Erstellung der Testversionen nicht mehr separat auf die Schwierigkeit geachtet werden musste. Weiters wurde am LUCET durch luxemburgischsprachige Mitarbeiter getestet, wie viele Items in einem vertretbaren Zeitrahmen seriös und gewissenhaft bearbeitet werden konnten. Die Ergebnisse sprachen für eine obere Grenze von 150 Items/ 30 Testseiten bei einem Median von etwa 21 Minuten.

Als Steigerung der Idee des unvollständigen Blockdesigns, wurde als nächster Schritt angenommen, dass jeder Teilnehmer einen individuellen Test bearbeitet, der nur über wenige Itemblöcke mit anderen Testversionen verbunden ist. Auf Basis der 207 Testseiten wurden insgesamt 1000 verschiedene Testversionen algorithmisch erstellt, wobei a) jede Testseite

insgesamt mindestens 50 mal und b) möglichst gleich häufig mit allen anderen Testseiten zusammen vorkommen sollte (Verlinkung). Dazu wurden für jeden Test nacheinander zufällig jeweils 30 Testseiten gezogen, wobei die Wahrscheinlichkeit einer Seite, in den aktuellen Test integriert zu werden, abhing von:

- Ihrer Häufigkeit in bisherigen Tests: Eine höhere Wahrscheinlichkeit gab es, wenn sie bereits mindestens einmal verwendet wurde, aber noch nicht 50 Mal.
- Der Verbesserung der "Verlinkungs-Statistik". Seiten, die mit den bisher im gerade entstehenden Test befindlichen Seiten noch wenig verlinkt waren, erhielten eine höhere Wahrscheinlichkeit, gezogen zu werden.

Dabei zeigte sich, dass ab etwa 400 Versionen beinahe alle Items getestet werden können (Abbildung 1, links). Die Zuteilung der 5er-Blöcke durch den Algorithmus erfolgte ab etwa Testindex 400 nur aufgrund der Verlinkungs-Statistik, da die Minimalkriterien für die Seitenhäufigkeit erreicht wurden. Da aufgrund von Erfahrungswerten am LUCET ab etwa 50 Getesteten von robusten Schwierigkeitseinschätzungen ausgegangen werden kann, zeigte sich in der Simulation, dass mit dem geplanten Ansatz bereits eine Stichprobengröße von 300 Teilnehmern für die Kalibrierung des Großteils der Items ausreichen würde (Abbildung 1, rechts).

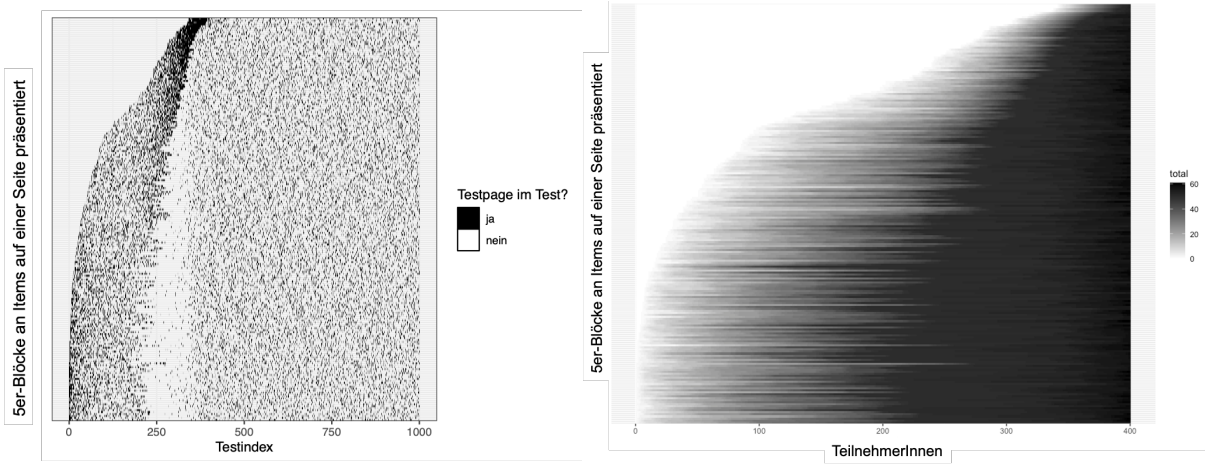

Abbildung 1, links: Verteilung der einzelnen Testseiten (Testpages) über die einzelnen Testversionen (Testindex); rechts: kumulierte Häufigkeit der Testseiten über die ersten 400 Testversionen

Um diesen Ansatz zur Verlinkung der einzelnen Versionen zusätzlich zu testen, wurde eine Simulationsstudie durchgeführt. Damit sollte sichergestellt werden, dass die Item- und Personenparameter mit zufriedenstellender Genauigkeit geschätzt werden können. Wie auch für die spätere Kalibrierung berücksichtigt, wurde dabei ein eindimensionales Rasch-Modell (siehe z.B. Fischer & Molenaar, 1995) als Grundlage herangezogen. Das Rasch-Modell beschreibt das Antwortverhalten der Getesteten als probabilistische Funktion und geht davon aus, dass der Leistung in einem Test eine einzige Fähigkeit bzw. Kompetenz zugrunde liegt. Die einzelnen Aufgaben sind voneinander unabhängig und die Anzahl der als korrekt beantworteten Items dient als ausreichende Statistik für die Schätzung der Personenfähigkeit. Verschiedene Modelltests bieten eine Einschätzung, inwieweit ein Test dem Modell entspricht.

Der Simulationsstudie lagen nun die Einschätzungen der Reviewer zur Aufgabenschwierigkeit zugrunde: Für als einfach eingeschätzte Aufgaben, wurde eine mittlere Lösungswahrscheinlichkeit von 90% vorgesehen, für mittelschwere Items 70% und für schwere Items 50%. Diese Einstufung wurde seitens ZLS bestätigt.

Unter diesen Prämissen wurden Itemparameter berechnet, unter der Annahme einer normalverteilten Fähigkeit und einer Zufallskomponente. Daraus wurde eine Antwortmatrix mit 400 Fällen simuliert. Diese enthielt aufgrund des Testdesigns geplant 86% fehlende Werte, was aber für die Parameterschätzung kein Problem darstellte. Die Korrelation der spezifizierten mit den geschätzten Itemparametern betrug r = 0,89, in allen Fällen schloss das 95%-Konfidenzintervall der Parameterschätzer den spezifizierten Wert mit ein (siehe Abbildung 2).

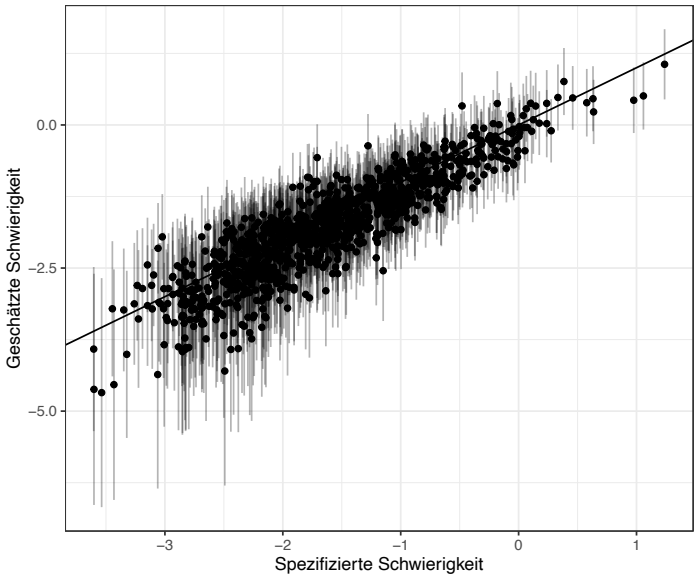

Abbildung 2: Zusammenhang vorab spezifizierter Itemschwierigkeiten mit durch die Simulationsstudie geschätzten Itemschwierigkeiten

Die Ergebnisse der Simulationsstudie waren insgesamt also sehr vielversprechend. So konnte gezeigt werden, dass selbst bei deutlich weniger Teilnehmern als geplant bzw. erwartet (bspw. 100 oder 200), oder Wegfall erwartbar problematischer Items, die Verlinkung der einzelnen Itemblöcke noch robust genug für stabile Itemparameterschätzungen für den Großteil der Items ist. Dies gab den endgültigen Ausschlag dafür, die 1000 für die Simulation erstellten Testversionen auch für die Pilotierung heranzuziehen. Diese Variante hatte den zusätzlichen Vorteil einer höheren Testsicherheit, da jeder Proband eine einzigartige Testversion bearbeiten würde.

Stichprobe und Studiendesign

Die Stichprobe zur Pilotierung der entwickelten Items sollte einerseits möglichst repräsentativ für die Zielpopulation sein, zum anderen aber auch groß genug, um möglichst alle Items ausreichend zu testen. Aufgrund der hohen Anzahl an entwickelten Items war die Bestimmung der optimalen Stichprobengröße nicht trivial. Zu diesem Zweck wurde eine Simulationsstudie zur Abschätzung der unteren Grenze der benötigten Teilnehmeranzahl zur robusten Abschätzung psychometrischer Parameter festgelegt (siehe oben).

Auf Basis dieser Einschätzung rekrutierte das ZLS landesweit Studienteilnehmer mit dem Ziel, eine repräsentative Stichprobe in Bezug auf folgende Aspekte zu gewinnen:

- Geschlecht
- Alter
- Bildungsgrad
- Die Regionen Luxemburgs
- Allgemeine Sprachkompetenz

Dies sollte zudem gewährleisten, dass die entwickelten Items auf ihre Fairness hinsichtlich wichtiger Kriterien geprüft werden konnten. So war vorab anzunehmen, dass es bei manchen geprüften Wörtern regionale oder altersbedingte Unterschiede der Aussprache und somit auch der angenommenen Schreibweise geben könnte.

Die Datenerhebung fand von 21.4.2022 bis inkl. 31.7.2022 statt, wobei insgesamt Daten von 705 Teilnehmern (56% weiblich, 2,5% divers oder keine Angabe) gesammelt werden konnten, die den Orthografietest als auch den anschließenden Fragebogen beantwortet haben. Mehrere Schulklassen konnten u.a. im Rahmen ihres Unterrichts für die Teilnahme gewonnen werden (Athénée de Luxembourg, Lycée Michel Rodange, Lycée Hubert Clément Esch, Lycée de Garçons Esch, École Sainte-Anne, Lycée Mathias Adam, Lycée Josy Barthel Mamer, International School Jonglinster). Zudem wurde an der Universität Luxemburg und am ZLS selbst getestet. Dazu nahmen Mitarbeiter des Institut national d'administration publique (INAP), des Compte rendu der Chambre des Députés und des Institut national des langues Luxembourg (INLL) an der Pilotierung teil. Weiters konnten zusätzliche Teilnehmer über einen Aufruf auf der Medienplattform RTL gewonnen werden. Die Testung fand dabei auf Tablets statt, unter Verwendung der Online-Testplattform OASYS des Luxembourg Centre for Educational Testing (LUCET).

Aus datenschutzrechtlichen Gründen wurde das Alter der Teilnehmer in vorab definierten Kategorien abgefragt. Etwa 56% gaben ein Alter unter 21 Jahren an, etwas mehr als ein Drittel befand sich im berufstätigen Alter zwischen 21 und 60 (37,1%, siehe Tabelle 4). Die in Hinblick auf den Bildungsgrad größten Gruppen der Stichprobe, hatten entweder (noch) keinen Bildungsabschluss (56%) oder ein Universitätsdiplom (28%), siehe Tabelle 5. Dies spiegelt sehr gut die hauptsächliche Rekrutierung an Schulen, als auch an staatlichen Institutionen wider.

Tabelle 4: Alter der Teilnehmer in abgefragten Kategorien

Alter	Häufigkeit	Prozent	Kumulative Prozent
21 - 30	72	10,2	10,2
31 - 40	72	10,2	20,4
41 - 50	62	8,8	29,2
51 - 60	56	7,9	37,2
61 - 70	24	3,4	40,6
71 - 80	14	1,9	42,6
< 21	394	55,9	98,4
> 80	4	0,6	99,0
Keng Angab	7	0,9	100
Missing	0	0,0	
Total	**705**	100	

Tabelle 5: Höchster Bildungsabschluss in abgefragten Kategorien

Ausbildungsofschloss	Häufigkeit	Prozent	Kumulative Prozent
(Nach) keen Ofschloss	397	56,3	56,3
Beruffsausbildung (Diplôme de technicien, Diplôme d'aptitude professionnelle (fréier: CATP), Certificat de capacité professionnelle (fréier: CITP oder CCM) oder e vergläichbaren auslänneschen Ofschloss)	30	4,3	60,6
Héichschouldiplom (BTS, fréier: ISERP, IEES, IST)	24	3,4	63,9
Meeschterbréif (Brevet de maîtrise)	11	1,6	65,5
Premièresdiplom (Diplôme de fin d'études secondaires, Bac, Baccalauréat international, A-levels)	48	6,8	72,3
Universitéitsdiplom (Bachelor, Master, Doktorat)	195	27,7	100
Missing	0	0,0	
Total	**705**	100	

Messinstrumente

Orthografietest

Der Orthografietest wurde mittels der online Testplattform OASYS auf Tablets im Kioskmodus administriert, dies verhinderte, dass andere Websiten während der Testung aufgerufen werden konnten. Jeder Getestete bearbeitete einen vorab erstellten, individuellen Test mit insgesamt 150 Items. Diese wurden jeweils in Kategorie-übergreifenden 5er-Blöcken pro Seite präsentiert. Um Lerneffekte durch einzelne Items zu verhindern, konnten die Teilnehmer dabei nicht zu bereits bearbeiteten Testseiten zurückspringen. Jedes Item bestand aus einem Itemstamm - einem Satz mit Wortlücke - und 4 Multiple-choice Antworten (eine Lösung, drei Distraktoren). Die Position der Lösung wurde dabei randomisiert, sodass der Einfluss von response sets weitgehend minimiert werden konnte. Die zur Verlinkung der unterschiedlichen Testversionen verwendeten 5er-Blöcke bzw. Testseiten wurden dabei an unterschiedlichen Positionen im Test präsentiert um Reihenfolgeeffekte auf die geschätzten Itemparameter weitgehend zu vermeiden.

Demografischer Hintergrund der Teilnehmer

Zur Bestimmung der Herkunft wurde die Heimatgemeinde der Teilnehmer abgefragt. Hierzu wurde eine Liste von 102 luxemburgischen Gemeinden zur Auswahl angeboten. Für die Auswertungen wurden jeweils Kantone zusammengefasst, um die Regionen Norden (Clervaux, Wiltz, Diekirch, Vianden), Zentrum (Redange, Mersch, Capellen, Luxemburg), Süden (Esch-sur-Alzette) und Osten (Echternach, Grevenmacher, Remich) möglichst adäquat zu repräsentieren. Anhand dieser Einteilung wurde geprüft, inwieweit Items in den einzelnen Regionen psychometrisch gleich funktionieren (=Fairness).

Testakzeptanz

Zur Abschätzung der Testakzeptanz wurden 14 Items des publizierten Fragebogens Akzept (Kersting, 1998) adaptiert und übernommen. Es sollten so Einblicke gewonnen werden zur wahrgenommenen Messqualität des Orthografietests (4 Items), seiner Augenscheinvalidität (2 Items), der Beeinflussbarkeit des Testergebnisses (Kontrollierbarkeit, 4 Items), als auch zu welchem Grad der Test als belastend erlebt wird (Belastungsfreiheit, 4 Items). Die 14 Items bestanden aus Einzelaussagen, die die Testteilnehmer auf einer 6-stufigen Likertskala hinsichtlich Richtigkeit beurteilen mussten (1 = trifft nicht zu, 6 = trifft voll zu). Die eingesetzte Fragebogenversion zeigte hohe Reliabilität (Cronbach's alpha = 0,87) und konnte gut die einzelnen Aspekte differenzieren.

Weiters gab es für die Teilnehmer die Möglichkeit, in einem offenen Kommentarfeld noch Rückmeldung und Feedback zum Test zu geben.

Ergebnisse

Eindimensionalität und Schwierigkeit der Items

Alle Items (n = 1035) wurden zwischen 97 und 106 mal bearbeitet. Generell fiel der Test bei sehr hoher Messgenauigkeit (Cronbach's alpha = 0,95) über alle Fähigkeitsbereiche hinweg, sehr leicht aus (siehe Abbildung 3), was vermutlich auf eine Verzerrung aufgrund der Zusammensetzung der Stichprobe zusammenhing (überdurchschnittlich hohe Bildungsabschlüsse). Dies führte u.a. allerdings dazu, dass manche Items kaum mehr zwischen Teilnehmern mit niedrigen und hohen Orthografiekenntnissen differenzieren konnten (Korrelation zwischen Lösung des Items und Gesamtscore im Test, $r < 0,25$). Aufgrund dieser niedrigen Trennschärfe mussten 171 Items ausgeschieden werden.

Eine daran anschließende Principal Component Analyse (PCA) der standardisierten Residuen zeigte eine „ausreichende" Eindimensionalität (Chou & Wang, 2010), das heißt, dass die Leistungen im Test vor allem durch einen statistischen Faktor (einer „Fähigkeit") erklärt werden können. Der stärkste Faktor der PCA erklärte lediglich 5% der Residualvarianz und es konnte keine inhaltliche Gemeinsamkeit der Items erkannt werden. Somit ist die Grundvoraussetzung des Rasch Modells erfüllt (Fischer & Molenaar, 1995), dass der Summenscore des Tests eine adäquate Abbildung der Kompetenzunterschiede der Orthografiekenntnisse der Teilnehmer ist.

Zur detaillierten Analyse der einzelnen Distraktoren pro Item, wurde für jede Antwort pro Item der durchschnittliche Gesamtscore für alle restlichen bearbeiteten Items gebildet. Erwartungsgemäß müsste dieser Score für korrekte Antworten höher sein als für falsche Antworten, da davon auszugehen ist, dass fähigere Teilnehmer eher die Lösung ankreuzen. Items, bei denen einzelne Distraktoren einen derart höheren Gesamtscore hatten (i.e. die Teilnehmer, die diesen Distraktor gewählt hatten, schnitten im restlichen Test besser ab, als die Teilnehmer welche die Lösung angekreuzt hatten) wurden markiert (73 Items) und gemeinsam mit dem ZLS auf fehlerhafte Distraktoren geprüft.

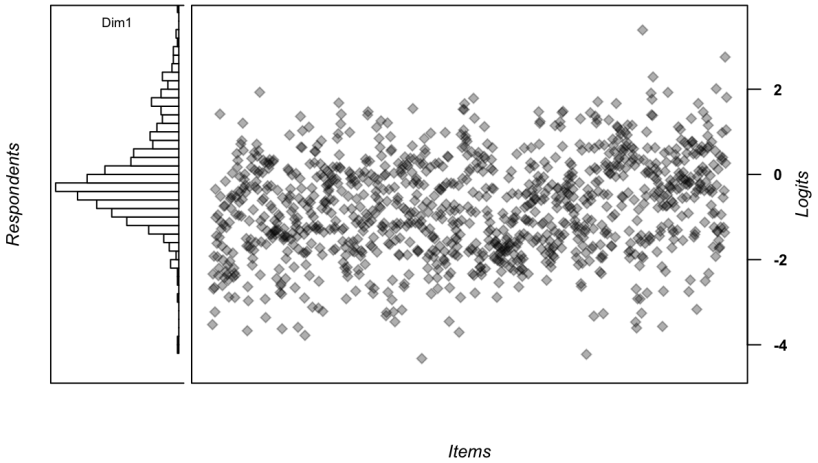

Abbildung 3: Gegenüberstellung von Personenfähigkeit und Itemschwierigkeit (Wrightmap)

Bemerkenswert ist, dass die Schwierigkeitseinschätzung der Reviewer am ZLS ziemlich nahe an der tatsächlichen empirischen Schwierigkeit lagen (siehe Abbildung 4, links). Des Weiteren lässt sich aussagen, dass einzelne Hauptkategorien der Orthografieregeln schwieriger sind als andere (z.B.: Groß- und Kleinschreibung, Getrennt-/ Zusammenschreibung und generell Fremdwörter (Abbildung 4, rechts).

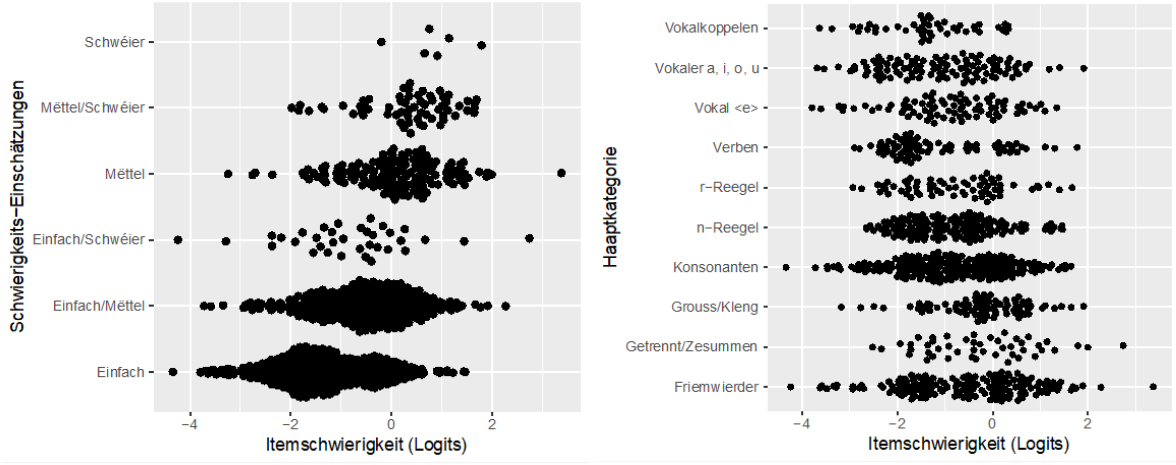

Abbildung 4: Itemschwierigkeiten je nach eingeschätzter Schwierigkeit durch Reviewer (links); Itemschwierigkeiten je nach inhaltlicher Hauptkategorie (rechts)

Fairness der Items

Zur Analyse der Fairness der Items wurde untersucht, inwieweit Items für unterschiedliche Personengruppen, unterschiedliche Schwierigkeiten aufweisen. In diesem Fall würde man von Differential Item Functioning sprechen (DIF; siehe z.B. Osterlind & Everson, 2009): eine oder mehrere Gruppen hätten einen systematischen Vor- oder Nachteil. Untersucht wurde, ob es Unterschiede hinsichtlich Alter (Personen <= 21 Jahre vs. Personen > 21 Jahren), Geschlecht (Männer vs. Frauen) und Herkunft (Zentrum vs. Rest). Insgesamt wiesen 65 Items statistisch signifikante, bedeutsame (Koeffizient in dem berechneten Facetten Modell > 2) Gruppenunterschiede auf.

Testdauer und-akzeptanz

Zur besseren Abschätzung der Bearbeitungszeit wurden den Teilnehmern keine Einschränkungen gesetzt. Wie in Abbildung 5 ersichtlich, war die große Mehrheit nach 40 Minuten Testzeit fertig. Gleichzeitig weist die Verteilung auf eine sehr hohe Plausibilität der gewonnenen Daten hin.

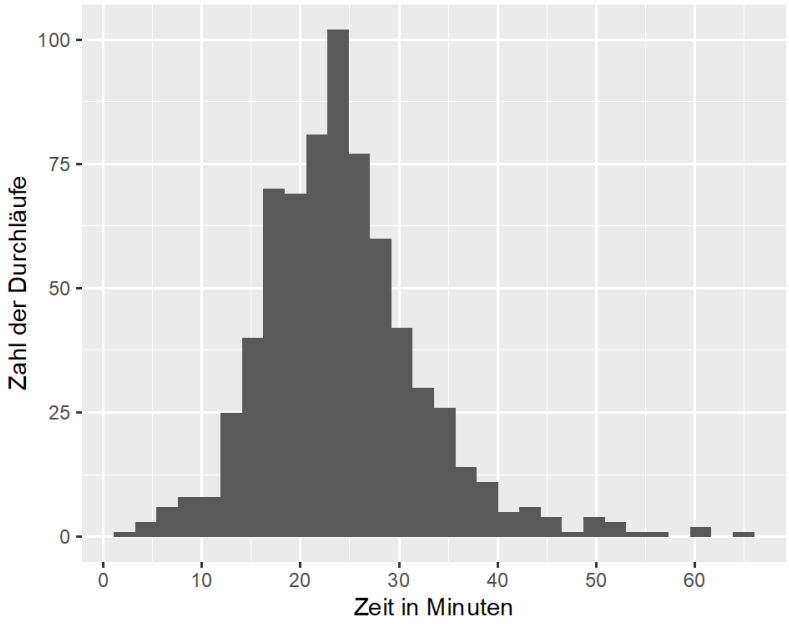

Abbildung 5: Anzahl der Testdurchläufe per Testdauer in Minuten

Dem Orthografietest wird eine solide Messqualität bescheinigt (Mittelwert = 4,00, siehe Tabelle 6), die große Mehrheit der Teilnehmer sind also der Meinung, dass der Test zuverlässig Leistungsunterschiede hinsichtlich der Rechtschreibleistung abbildet. Dies findet sich in der attestierten Kontrollfähigkeit wieder (Mittelwert = 3,58), die Testaufgaben wurden großteils klar und verständlich empfunden, die Aufgabenstellung war klar und die Teilnehmer wussten, was zu tun ist. Mittlere Zustimmungswerte wurden für die Aspekte Augenscheinvalidität und Belastungsfreiheit gegeben. Die vom Test gestellten Aufgaben werden als ausreichend alltagsnah und nicht zu anstrengend erlebt.

Auch in den offenen Kommentaren kamen überwiegend positive Rückmeldungen und Lob für den Test. Einzelne kritische Kommentare oder Verständnisschwierigkeiten einzelner Items wurden aufgenommen und in weiterer Folge für den Itempool berücksichtigt.

Tabelle 6: Deskriptivstatistiken der administrierten Akzeptanzskalen

	Gültig	Fehlend	Mittelwert	SD	Min	Max
Kontrollierbarkeit	667	38	3,58	0,61	0	6
Messqualität	660	45	4,00	0,70	1	6
Augenscheinvalidität	665	40	3,13	0,79	0	6
Belastungsfreiheit	660	45	3,09	0,99	0	6

Fazit und finaler Itempool

Die Pilotierung kann im Allgemeinen als Erfolg gesehen werden, sowohl was die Testdurchführung, die psychometrische Qualität der eingesetzten Aufgaben, oder auch die Akzeptanz des Tests durch die Teilnehmer betrifft. Aufgrund der bereits oben erwähnten Kriterien (Trennschärfe, Distraktorenanalyse und DIF) wurden problematische Items aus dem Itempool ausgeschieden. Die meisten Items (171) scheiden aufgrund zu geringer Trennschärfe (< 0,25) aus, gefolgt von zu geringer Trennschärfe gepaart mit DIF in einer der Subgruppen (21). Tabelle 7 gibt einen genauen Überblick über die Gründe des Ausscheidens von Items. Der Anteil von etwa 24% ausgeschiedenen Items ist typisch für die Entwicklung von Leistungstests, kann sogar als eher niedrig betrachtet werden angesichts der umfangreichen Qualitätsanalysen. Einmal mehr zeigt sich, dass sich die mit relativ hohem Aufwand betriebene Itementwicklung absolut gelohnt hat.

Insgesamt stand nach der Anwendung diverser Kriterien also ein finaler Itempool von 787 psychometrisch kalibrierten und fairen Aufgaben zur Verfügung, deren Verteilung auf die getesteten Hauptkategorien in etwa der Verteilung aller entwickelten und im Test vertretenen Items entspricht.

Tabelle 7: Anzahl ausgeschiedener Items nach Kriterium

Kriterium	**Anzahl ausgeschiedener Items**
Niedrige Trennschärfe (< .25)	171
Niedrige Trennschärfe + DIF	21
DIF + > 95% korrekte Antworten in einer Subgruppe	19
Niedrige Trennschärfe + DIF + >95% korrekte Antworten in einer Subgruppe	12
DIF	10
Niedrige Trennschärfe + problematischer Distraktor	8
DIF + Niedrige Trennschärfe + problematischer Distraktor	3
> 95% korrekte Antworten in einer Subgruppe	2
Niedrige Trennschärfe + >95% korrekte Antworten in einer Subgruppe	2
Total	**248**

Validierungsstudie

Zielsetzung

Die Pilotierung lieferte zusätzlich zur psychometrischen Evaluation der Items auch zahlreiche Hinweise für eine optimale Testadministration für den finalen Einsatz des Orthografietests (Schwierigkeit, Testdauer, etc.). Nun sollten im nächsten und abschließenden Entwicklungsschritt, einerseits die geplante Form der Testadministration, andererseits aber auch die Güte des Tests im Vergleich zu bisher eingesetzten Verfahren getestet werden. Zu diesem Zweck wurde im Frühsommer 2023 eine Studie zur Kriteriumsvalidität durchgeführt. Als „Goldstandard" und Vergleichskriterium diente dabei ein üblicherweise zur Zertifizierung luxemburgischer Orthografiekenntnisse eingesetztes, standardisiertes Diktat.

Methode

Stichprobe und Studiendesign

Ähnlich zur Pilotstudie wurde die Rekrutierung der Teilnehmer vom ZLS organisiert. Die Stichprobe sollte dabei möglichst repräsentativ für die geplanten Anwendungsfälle des Orthografietests sein. Der Aufruf zur freiwilligen Teilnahme erfolgte daher wieder an Weiterbildungsinstituten, staatlichen Institutionen, als auch letztlich öffentlich (https://www.rtl.lu/kultur/news/a/2081085.html). Getestet wurde erneut am ZLS selbst, an der Universität Luxemburg, beim Compte rendu der Chambre des Députés https://www.chd.lu/fr/chamberblietchen und am INLL https://www.inll.lu/fr/. Unter anderem wurde am INLL der Test für den Semesterabschluss eines Lehrganges eingesetzt. Die Moien ASBL https://moienasbl.lu/ nahm im Rahmen einer Weiterbildung am Test teil. Außerdem konnte der Test für das Ifen-Lehrpersonal https://ssl.education.lu/ifen/ im Rahmen einer Weiterbildung angeboten werden. Zwei Schulklassen der IVe wurden getestet, die in diesem Jahr Orthografie als Unterrichtsfach hatten und deren Lehrer und Schüler am Test teilnehmen wollten. Da die Testung anonym erfolgte und keine weitere Erhebung demografischer Merkmale der Teilnehmer vorgesehen war, sollte dieser breite Aufruf erneut eine sehr heterogene Stichprobe ermöglichen (Fortbildner, Rentner, Jugendliche, Nicht-Luxemburgische Staatsbürger, Sprachbegeisterte, Teilnehmer mit guten Kenntnissen, oder auch lediglich Interessierte, ohne Vorbereitung oder ohne weitere Kenntnisse der Orthografie. Insgesamt fand die Datenerhebung von 2. Mai 2023 bis 15. Juli 2023 statt. In diesem Zeitraum konnten 176 Personen vollständig getestet werden.

Die Testsessions dauerten dabei jeweils eine Stunde in der zuerst ein standardisiertes Diktat und danach der Orthografietest unter Aufsicht von einer oder zwei Personen durchgeführt wurden. Die Teilnehmer bekamen Feedback zu ihrer Leistung in Form von Anzahl korrekter Antworten. Getestet wurde auf 32 iPads, die vom CGIE https://portal.education.lu/cgie/ so konfiguriert waren, dass man nur auf die Testplattform OASYS (https://cbt.lucet.lu/zls/) zugreifen konnte. Am INLL wurde mit vorhandener Infrastruktur ohne Probleme getestet.

Messinstrumente

Orthografietest

Die 787 psychometrisch kalibrierten Items aus der Pilotierungsstudie wurden erneut vom ZLS kritisch durchgesehen. Dabei wurde besonderes Augenmerk auf Items gelegt, zu denen a) es aus der Pilotierungsstudie Feedback zu Verständnisschwierigkeiten oder Missverständnisse der Teilnehmer gab, b) für deren Schreibweise eine Toleranzvariante existiert oder c) deren zugrundeliegende Orthografieregel nicht mehr getestet, in naher Zukunft eventuell geändert oder abgeschafft werden könnte. Da der Test für die nächsten Jahre möglichst ohne Änderungen laufen sollte, wurden daher schon für die Validierungsstudie insgesamt 68 Items gelöscht. Sieben Items wurden einer anderen, passenderen Kategorie zugeordnet. Außerdem wurden kleinere Änderungen am Itemstamm wie bspw. Tippfehler durchgeführt. Der finale Itempool von 719 Aufgaben, der für die Erstellung der Testversionen der Validierungsstudie zur Verfügung stand, findet sich in Tabelle 8.

Tabelle 8: Anzahl entwickelter Items pro Hauptkategorie

Hauptkategorie	**Häufigkeit**	**Prozent**
D'Vokaler a, i, o an u	87	12,2
De Vokal e	59 (61*)	8,26
D'Konsonanten	146 (147*)	20,4
D'Verben	57	7,98
D'Friemwierder	106	14,85
D'n-Reegel	103 (104*)	14,4
D'r-Reegel	49	6,8
Vokalkoppelen	40	5,6
D'Grouss- a Klengschreiwung	46	6,4
D'Getrennt- an Zesummeschreiwung	21 (22*)	2,9
Exoten	n.a.**	n.a.**
Total	**714 (719*)**	**100**

*Verwendete Anzahl an Items für Validierungsstudie, einzelne Items wurden aufgrund Feedback vom ZLS danach noch gelöscht
**Im Orthografietest wurden Items der Kategorie „Exoten" auf die anderen Kategorien aufgeteilt

Auf Basis der positiven Ergebnisse der Pilotierungsstudie, wurde für die Validierung (und dementsprechend auch für die geplante finale Form des Orthografietests) erneut auf den Ansatz individueller Testversionen zurückgegriffen. Zusätzlich wurde bei deren Erstellung berücksichtigt, dass Wörter, die in einem Itemstamm vorkommen, in keinem Item derselben Testversion vorkommen. Eine zusätzliche Kontrollebene wurde eingeführt, die sicherstellte, dass Items, die dasselbe Wort abfragten, ebenfalls nicht mehr im selben Test sind. Die ersten beiden Testpages wurden absichtlich leichter gestaltet, um den Einstieg in den Test einfacher und angenehmer zu machen. Danach variierte die Schwierigkeit der 5er-Blöcke pro Testseite wieder, ganz ähnlich zur Pilotierungsstudie.

Die Auswahl der Items für die einzelnen Testversionen erfolgte aufgrund ihrer psychometrischen Kennwerte und der gemessenen Hauptkategorie. Um den Itempool bestmöglich auszuschöpfen, folgte die Verteilung auf die einzelnen Kategorien einem Mittelweg zwischen ursprünglich angedachter Gewichtung und Anzahl verfügbarer Items. Dies entspricht einerseits der vereinbarten inhaltlichen Abdeckung (construct coverage) des Konstrukts Orthografiekenntnisse, garantierte aber andererseits, dass möglichst alle kalibrierten Items zum Einsatz kamen und so die Testsicherheit erhöht wurde. Wie oft ein Item ausgewählt wurde, hing v.a. davon ab, wie der Anteil der Kategorien im Gesamtpool im Verhältnis zur gewünschten Verteilung stand. Waren Kategorien im Pool überrepräsentiert, wurden die einzelnen Items dieser Kategorien natürlich auch im Verhältnis seltener gezogen. Aufgrund der gut argumentierbaren Eindimensionalität des gesamten Itempools, war die Orientierung an den Itemhäufigkeiten pro Kategorie aber durchaus gerechtfertigt und eine vergleichbare Messung garantiert, selbst wenn die inhaltliche Verteilung pro Testversion leicht abweicht.

Weiters wurde bei der Itemauswahl sichergestellt, dass die Testversionen sich hinsichtlich der (Item-)Schwierigkeiten möglichst gleichen (operationalisiert anhand Verteilung, Mittelwert und Median) und möglichst günstige Kennwerte für DIF-Werte hinsichtlich Geschlecht, Alter und Region sowie Trennschärfe aufweisen.

Um nun die besten Testversionen für den Anwendungsfall zu ermitteln, wurden nach obigem Prinzip zufällig mehrere Millionen Testversionen erstellt und folgende Kennwerte pro Version berechnet:

- Eine Kolmogorov-Smirnov Statistik (KSD), die die Abweichung der Schwierigkeitsverteilung der Version zu der gemittelten Schwierigkeitsverteilung von zehntausend zufällig, nach der empirischen Kategorienverteilung erstellter Tests abbildet
- Mittlere Schwierigkeit
- Median der Schwierigkeit
- Mittlerer DIF hinsichtlich Alter, Geschlecht und Region
- Mittlere Trennschärfe

Diese Kennwerte wurden zu einem Testqualitäts-Indikator verrechnet, nach dem alle generierten Testversionen sortiert wurden. Die dabei besten 100 Testversionen wurden letztlich für die Validierungsstudie herangezogen. Durch eine weitere Simulationsstudie wurde sichergestellt, dass die Fähigkeit der Testteilnehmer unabhängig von der ihnen zugewiesenen Testversion korrekt geschätzt werden kann, und dies über das gesamte Fähigkeitsspektrum.

Aufgrund der Erfahrungen aus der Pilotierungsstudie, wurde die Testlänge auf 90 Items pro Testversion (also 18 Testseiten) reduziert. Entsprechende Berechnungen zeigten, dass für die Messgenauigkeit mit keinen Einbußen gerechnet werden musste. Für die Bearbeitung der Testversion hatten die Teilnehmer maximal 40 Minuten Zeit. Um Lerneffekte während der Testbearbeitung zu verhindern, war es ähnlich zur Pilotierung wieder nicht möglich, bereits bearbeitete Itemblöcke aufzurufen.

Diktat

Bisher kamen zur Feststellung der luxemburgischen Orthografiekenntnisse standardisierte Diktate zum Einsatz. Dieser „Goldstandard" sollte als Kriterium zur Validierung des Orthografietests dienen und gleichzeitig einen Vergleich ökonomischer Gesichtspunkte erlauben. Dafür wurden 28 kurze Sätze eingesprochen und für die Testung standardisiert abgespielt. Dabei wurde jeder Satz dreimal wiederholt. Die Testteilnehmer mussten dazu auf Papier einen Lückentext mit insgesamt 58 fehlenden Wörtern ausfüllen. Um eine vergleichbare Schwierigkeit zum Orthografietest zu erreichen, wurde die prozentuale Verteilung der Wortlücken (= Items) auf die Hauptkategorien möglichst identisch gestaltet (siehe Tabelle 9). Die Testung dauerte etwa 11 Minuten. Die Reliabilität des Diktats lag bei Cronbach's alpha = 0,97.

Tabelle 9: (Prozentualer) Anzahl an Items per Kompetenzbereich für Diktat und Orthografietest

	Diktat		**Orthografietest**	
	Anzahl Items	**Prozentualer Anteil (Basis 58 Wörter)**	**Durchschnittliche* Anzahl Items**	**Prozentualer Anteil (Basis 90 Items)**
Kompetenzbereich				
D'Vokaler a, i, o an u	8	13,8	13	14,4
De Vokal e	8	13,8	13	14,4
D'Konsonanten	8	13,8	13	14,4
D'Verben	5	8,6	9	10
D'Friemwierder	5	8,6	9	10
D'n-Reegel	5	8,6	9	10
D'r-Reegel	5	8,6	9	10
Vokalkoppelen	5	8,6	9	10
D'Grouss- a Klengschreiwung	3	5,1	3	3,3
D'Getrennt- an Zesummeschreiwung	3	5,1	3	3,3
Exoten	3	5,1	n.a**	n.a**
Summe	**58**	**100**	**90**	**100**

*Tatsächliche Anzahl der Items kann maximal um 1 abweichen
**Im Orthografietest wurden Items der Kategorie „Exoten" auf die anderen Kategorien aufgeteilt

Ergebnisse

Deskriptivstatistik

Zur besseren Übersicht der Leistungen in Diktat als auch Orthografietest, sind die Deskriptivstatistiken in Tabelle 10 wiedergegeben. Zusätzlich zu den Summenwerten in beiden Tests, wurden diese in Prozent der maximal erreichbaren Punkte umgerechnet, können damit also unmittelbar verglichen werden.

Es fällt direkt auf, dass beide Tests annähernd gleich schwierig sind, der Median für das Diktat bei 77,6 erreichten Prozent, für den Orthografietest bei 81,1 erreichten Prozent liegt. Beide Verfahren decken den gesamten Fähigkeitsbereich ab, mit einem leichten Deckeneffekt. Die Tests differenzieren vor allem im mittleren Fähigkeitsbereich sehr gut.

Tabelle 10: Deskriptivstatistiken für Diktat und Orthografietest

	Summe Diktat	Summe Orthografietest	Prozent Diktat	Prozent Orthografietest
Gültig	176	176	176	176
Median	45	73	77,6	81,1
Mittelwert	39,54	68,35	68,2	75,9
Standardabweichung	14,78	17,04	25,5	18,9
Minimum	7	29	12,1	32,2
Maximum	58	90	100	100

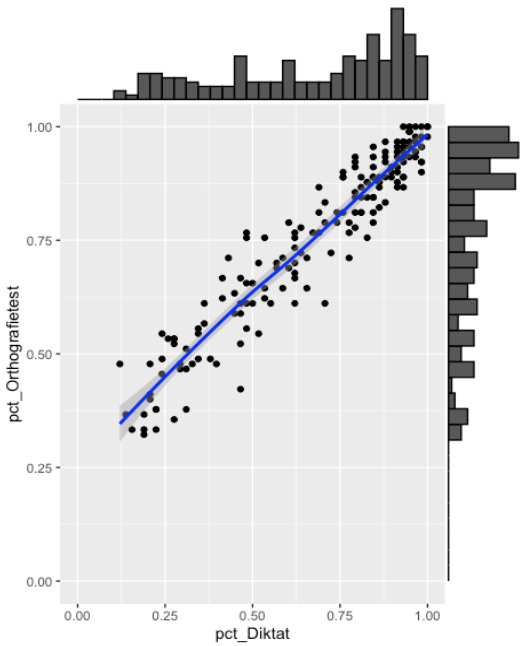

Abbildung 6: Histogramm und Streudiagramm der Testergebnisse aus Orthografietest (y-Achse) und Diktat (x-Achse). Anmerkung: die in blau dargestellte Linie im Plot ist ein LOESS-Smoother und zeigt die Linearität des Zusammenhangs

Kriteriumsvalidität

Aufgrund des metrischen Skalenniveaus des Diktats (Anzahl gelöster Aufgaben), als auch des Orthografietests (Prozentsatz gelöster Aufgaben) und des in Abbildung 6 ersichtlichen, annähernd linearen Zusammenhangs, wurde zur Bestimmung der Kriteriumsvalidität eine Pearson Korrelation berechnet. Die beobachtete Korrelation von $r = 0{,}95$ ($p < .05$) lässt auf einen sehr hohen Zusammenhang beider Leistungen schließen und spricht daher eindeutig für die Kriteriumsvalidität des entwickelten Orthografietests. Aufgrund der höheren Itemanzahl des Orthografietests, ist auch davon auszugehen, dass dieser im Vergleich zum Diktat genauer misst.

Fazit

Nach erfolgreicher Pilotierung und psychometrischer Kalibrierung der entwickelten Items stand ein finaler Pool von 719 Aufgaben zur Verfügung. Daraus wurden 100 Testversionen mit vergleichbaren psychometrischen Eigenschaften erstellt. Diese sollten anhand eines etablierten Verfahrens validiert werden. Anhand der durchgeführten Validierungsstudie konnte eindrucksvoll die Äquivalenz des Orthografietests zum bisher eingesetzten Diktat demonstriert werden. Sowohl die annähernd identische Abbildung der Fähigkeitsverteilung als auch die sehr hohe Korrelation sprechen für die hohe Qualität des entwickelten Orthografietests.

Dem ZLS steht mit dem in OASYS implementierten Orthografietest somit ein qualitativ hochwertiges Messinstrument zur Verfügung, um valide und reliabel die luxemburgischen Rechtschreibkenntnisse in einem sehr ökonomischen Zeitrahmen zu testen.

Referenzen

Chou, Y. T., & Wang, W. C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. Educational and Psychological Measurement, 70(5), 717-731.

Conseil fir d'Lëtzebuerger Sprooch (CPLL) & Zenter fir d'Lëtzebuerger Sprooch (ZLS). (2021) D'Lëtzebuerger Orthografie. Luxembourg: SCRIPT & ZLS.

Downing, S. M., & Haladyna, T. M. (2006). Handbook of Test Development. Mahwah, N.J.: Lawrence Erlbaum Associates.

Fischer, G. H., & Molenaar, W. (1995). Rasch Models. Foundations, Recent Developments, and Applications. Springer, New York.

Kersting, M. (1998). Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren. Zeitschrift für Arbeits- und Organisationspsychologie, 42, 61-75.

Osterlind, S. J., & Everson, H. T. (2009). Differential item functioning. Sage Publications.

Danksagung

An dieser Stelle möchten wir unseren besonderen Dank aussprechen. Er gilt allen engagierten und motivierten Teilnehmern der Pilotierungs- und Validierungsstudie, deren Beitrag entscheidend zum Erfolg dieses Projekts beigetragen hat. Unser herzlicher Dank gilt auch den Mitarbeitern des ZLS, der Leitung des INLL sowie dem Kommissar für die luxemburgische Sprache. Nicht zuletzt möchten wir den beteiligten Mitarbeitern am LUCET unseren tiefen Dank aussprechen. Ohne die gemeinsame Anstrengung und Hingabe von allen wäre dieses Projekt niemals möglich gewesen.