Nili hat Geburtstag: Dynamic Assessment im Bereich des Wortlernens für Kindergartenkinder*

Nili's birthday: A dynamic assessment of word learning for preschool children

Christoph Till, Julia Winkes

Zusammenfassung:

Hintergrund: Die Wortschatzdiagnostik bei sukzessiv mehrsprachigen Kindern ist erschwert, da klassische Tests das vorhandene Sprachwissen nur ungenau erfassen. Dynamic Assessment (DA) zielt darauf ab, stattdessen das Sprachlernpotenzial zu erfassen.
Zielsetzung: Untersucht wurden die Interrater-Reliabilität eines neuen DA-Verfahrens sowie seine Fähigkeit, zwischen Kindern mit logopädischem und ohne logopädischen Förderbedarf zu unterscheiden.
Methodik: Das Verfahren kombiniert Test-Teach-Retest, Graduated Prompting und Modifiability Scales. Es wurde bei 39 Kindergartenkindern mit Deutsch als Zweitsprache angewendet. Die Auswertung umfasste Reliabilitätsanalysen, Gruppenvergleiche und ROC-Analysen.
Ergebnisse: Die Interrater-Reliabilität schwankt zwischen gut (rezeptive Wortschatzüberprüfung: α = .84) und gering (Modifiability Scales: α = .50). Signifikante Gruppenunterschiede zeigten sich im Gesamtscore (p = .014, d = –.76) und in den Modifiability Scales (p = .011). Die Sensitivität des Verfahrens lag bei 43–71 %, die Spezifität bei 64–96 %.
Schlussfolgerungen: Das Verfahren bildet kindliches Verhalten im Wortlernprozess ab, dessen Beurteilung jedoch mit Unsicherheiten behaftet ist. Es erlaubt zudem keine zuverlässige Unterscheidung zwischen Kindern mit und ohne Förderbedarf und misst das Sprachlernpotenzial möglicherweise nicht unabhängig vom bestehenden Wortschatzwissen. Zukünftige Entwicklungen sollten daher stärker untersuchen, welche Reaktionen in sprachbasierten Interaktionen tatsächlich verlässlich auf förderrelevantes Lernpotenzial hinweisen und wie diese sicher eingeschätzt werden können.

Schlüsselwörter

Sprachentwicklungsstörung; Deutsch als Zweitsprache; Wortschatz; Dynamic Assessment

Abstract

Background: The assessment of vocabulary in multilingual children presents a challenge, as conventional tests do not accurately reflect their language proficiency. Conversely, Dynamic Assessment (DA) endeavors to evaluate an individual's potential for language acquisition.
Aims: The interrater reliability of a novel DA procedure and its ability to differentiate between children with and without speech-language therapy needs was investigated.
Methods: The procedure combines Test-Teach-Retest, Graduated Prompting, and Modifiability Scales. Thirty-nine kindergarten children learning German as a second language were assessed. The evaluation included reliability analyses, group comparisons, and ROC analyses.
Results: Interrater reliability fluctuates between good (receptive vocabulary test: α = .84) and low (modifiability scales: α = .50). Group differences were found in the total score (p = .014, d = -.76) and in the modifiability scales (p = .011). Sensitivity ranged from 43 % to 71 %, specificity from 64 % to 96 %.

* Dieser Beitrag hat das double-blind Peer-Review-Verfahren durchlaufen.

Conclusions: The method depicts children's behavior in the word learning process, but its assessment is fraught with uncertainty. Its design may not permit independent measurement of language learning potential, as it seems contingent upon existing vocabulary knowledge. Future developments should investigate the link between word learning potential and existing vocabulary knowledge as well as the reliability of Modifiability Scales.

Keywords

Developmental Language Disorder; German as second language; vocabulary; dynamic assessment

1 Dynamic Assessment des Wortlernens von Kindern mit Deutsch als Zweitsprache

1.1 Wortschatzdiagnostik bei sukzessiv-bilingualen Kindern

Kinder, die im Alter von drei Jahren oder später eine neue Sprache erwerben, zählen zu der Gruppe der sukzessiv-bilingualen Kinder (Rothweiler & Ruberg, 2011). Während die Aussprache sowie die Grammatik im Laufe der Zeit mit zunehmender Sicherheit beherrscht werden, stellt der Erwerb eines altersgemäßen Wortschatzes eine besondere Herausforderung dar. So zeigen sukzessiv-bilinguale Kinder Schwierigkeiten im Bereich Wortschatz, die denjenigen von Kindern mit diagnostizierten semantisch-lexikalischen Störungen ähneln – v.a. in Bildbenennungsaufgaben erzielen sukzessiv-bilinguale Kinder im Schnitt nicht mehr Punkte als monolinguale Kinder mit einer SES (Sheng et al., 2006; Ehl et al., 2014; Wilkens et al., 2018).
Dies hängt mit der Besonderheit der Erwerbsaufgabe sukzessiv-bilingualer Kinder zusammen. So müssen diese nacheinander zwei Lexika aufbauen und beziehen den dafür nötigen Input aus mindestens zwei verschiedenen Quellen (Klassert & Kauschke, 2014), welche sich in puncto Quantität und Qualität deutlich unterscheiden können (Thordardottir, 2011). Vor allem vor der obligatorischen Schulzeit geschieht der Erwerb der Zweitsprache Deutsch eher unsystematisch, der Deutschwortschatz ist in den frühen Lebensjahren deutlich kleiner als in der Familiensprache. Erst ab Eintritt in die Schule und damit in den systematischen Deutschkontakt können bezüglich des Deutschwortschatzes deutliche Fortschritte beobachtet werden (Montanari et al., 2018; Lautenschläger et al., 2023). Dennoch bleibt der Wortschatz in der Umgebungssprache Deutsch von sukzessiv-bilingualen Kindern immer kleiner als der von monolingual deutschsprechenden Kindern. Während dieser sog. *vocabulary gap* (Thordardottir, 2011) im rezeptiven Wortschatz über die Jahre beinahe aufgeholt werden kann, gilt er im Bereich des expressiven Wortschatzes als stabil.

Beim Wortschatzaufbau spielt zudem der sog Matthäus-Effekt eine entscheidende Rolle: Je umfangreicher der vorhandene Wortschatz, desto leichter fällt es Kindern, neue Begriffe in ihr semantisch-lexikalisches Wissen zu integrieren (Juska-Bacher et al., 2024; Rupp, 2021). Auf Basis dieses Wissens können neue Wörter aus dem Kontext erschlossen, aufgrund der inhaltlichen Nähe zu bereits bekannten Wörtern hergeleitet und die Wortform wegen des vertrauten Klangs zu ähnlichen Wortformen schneller abgespeichert werden.
Dieser Lernprozess wird jedoch erschwert, wenn das semantisch lexikalische Wissen eher klein ist, keine verlässliche Grundlage bietet, und neue Wörter mit ihren verschiedenen Wortformen mühsam aus dem phonologischen Input herausgehört, mit Bedeutung versehen und in das bestehende Lexikon integriert werden müssen. Das ist z.B. dann der Fall, wenn die Kinder eine Sprachentwicklungsstörung (SES) haben, wodurch das sprachliche Entwicklungs- und Lernpotenzial eingeschränkt ist.

Aber auch bei Kindern mit einem hohen Sprachlernpotenzial können die Wortschatzleistungen eingeschränkt sein. Wenn die Kinder ein qualitativ und quantitativ eingeschränktes Sprachangebot in der Umgebung bekommen, können sie trotz bester Voraussetzungen ihr Sprachlernpotenzial nicht nutzen. In diesem Fall spricht man von umgebungsbedingten Sprachauffälligkeiten (Kauschke et al., 2023). Vor allem Kinder aus Familien mit sozio-ökonomischen Risikolagen sind betroffen, wobei Kinder mit Migrationshintergrund viermal häufiger zu dieser Gruppe gehören als Kinder ohne Migrationshintergrund (Rupp, 2021).
Um den besonderen Umständen des sukzessiv-bilingualen Worterwerbs gerecht zu werden, sollte man in der Wortschatzdiagnostik beide Sprachen des Kindes einbeziehen (als konzeptueller oder

als totaler Wortschatz gemessen). Nur dann kann das semantisch-lexikalische Wissen vollständig und damit gerechter beurteilt werden (Byers-Heinlein et al., 2024; Hoff et al., 2012; Montanari et al., 2018). Entsprechende Verfahren stehen jedoch nicht für jede Erstsprache zur Verfügung. Die diagnostische Erfassung der Wortschatzleistungen (wie auch anderer Sprachleistungen) sukzessiv-mehrsprachiger Kinder gilt vor diesem Hintergrund der genannten Zusammenhänge als problembehaftet (Holzinger et al., 2022). Fehldiagnosen bei sukzessiv-mehrsprachigen Kindern sind entsprechend häufig (Bonuck et al., 2022; Scherger, 2023). Freeman & Schroeder (2022) empfehlen deswegen in der Diagnostik alternative Methoden wie das Dynamic Assessment (DA).

1.2 Dynamic Assessment

Mit „klassischen" statusdiagnostischen Verfahren, bei denen Aufgabenformate mittels standardisierter Instruktionen möglichst ohne Beeinflussung durch die Testleitung durchgeführt werden, wird der aktuelle Lernstand bzw. das Sprachwissen einer Person erhoben. Dabei kann jedoch nicht differenziert werden, ob ein unterdurchschnittliches Ergebnis auf das eingeschränkte Lernpotenzial (im Sinne einer SES) der Person zurückzuführen ist oder seinen Ursprung in ungünstigen oder einfach anderen Lernbedingungen hat (Ehlert, 2021). Mittels DA hingegen soll das Lernpotenzial einer Person sichtbar werden, indem die Testleitung die Testperson bei der Bewältigung der Aufgaben unterstützt (Hasson & Joffe, 2007). Personen, die sich dank der Unterstützung unmittelbar verbessern können und somit ein hohes Lernpotenzial demonstrieren, sind folglich nicht auf eine intensive, therapeutische Unterstützung angewiesen. Diejenigen Personen jedoch, die trotz der Unterstützung durch die Testleitung ihre Testergebnisse nicht (wesentlich) verbessern können, zeigen ein geringes Lernpotenzial und sind daher auf intensive, langfristige, individuelle Unterstützung angewiesen, um die gewünschten Fortschritte erreichen zu können (Ehlert, 2021).

Eine bewährte Vorgehensweise wird *Test-Teach-Retest* genannt. Dabei wird ein (statusdiagnostischer) Test durchgeführt und das Ergebnis festgehalten. In der darauffolgenden Teach-Phase wird die getestete Fähigkeit gezielt trainiert. Im Anschluss an die Teach-Phase wird die Fähigkeit wiederum mit demselben Test oder mit einer Parallelversion überprüft. Im Vergleich der Punktzahl im Retest mit der des ersten Tests wird ersichtlich, ob und wieviel eine Person vom ersten zum zweiten Test hinzulernen konnte. Handelt es sich um einen bedeutenden Zuwachs, spricht dies für ein hohes Lernpotenzial und intensivere, therapeutische Maßnahmen sind nicht notwendig. Ist hingegen kein oder nur ein geringer Zuwachs festzustellen, steht dies für ein eingeschränktes Lernpotenzial. Im letzten Fall sind intensivere Unterstützungsmaßnahmen angezeigt, um die gewünschten Fortschritte erreichen zu können.

Bei der Entwicklung von Test-Teach-Retest-Verfahren gibt es jedoch einige Schwierigkeiten. So haben nicht alle Items des Tests die gleiche Itemschwierigkeit, d.h. dass manche Punkte in Tests schwieriger zu erreichen sind als andere und die Vergleichbarkeit der Zuwächse dadurch erschwert ist. Auch bereitet es Kindern mehr Mühe, Fortschritte zu machen, wenn sie schon nah an ihrem persönlichen Leistungsmaximum arbeiten. Kinder jedoch, die deutlich unter ihrem Potenzial agieren, haben „viel Luft nach oben" und können sich dementsprechend leichter und deutlicher steigern. Hinzukommt die Schwierigkeit, dass die Qualität der Teach-Phase sehr unterschiedlich ausfallen kann. Während es einer Person schnell gelingen mag, hilfreiche, zum jeweiligen Kind passende Instruktionen zu geben, schafft dies eine andere Person möglicherweise weniger gut. Der erzielte Erfolg in der Teach-Phase kann sich also auch in Abhängigkeit von der testenden Person deutlich unterscheiden. Dementsprechend ernüchternd sind daher auch Ergebnisse aus Metaanalysen und Reviews, die dem Test-Teach-Retest-Verfahren ungenügende diagnostische Leistungen bescheinigen (Orellana et al., 2019; Hunt et al., 2022).

Ein anderer Ansatz, *Graduated Prompting* genannt, zielt darauf ab, unmittelbar während der Testung Hilfestellungen anzubieten, damit die getestete Person die gestellten Aufgaben lösen kann. Die Hilfestellungen werden dabei hierarchisiert angeboten, d.h. von den am wenigsten hilfreichen zu den stärksten Hinweisen. Beim interventionist approach erfolgt das Geben von Hilfestellungen gemäß einem standardisierten Skript, sodass jede Person die gleichen Empfehlungen in der gleichen Reihenfolge erhält. Beim interactionist approach hingegen sind die Hilfestellungen individuell zu geben, sodass je nach Person unterschiedliche Hinweise erfolgen (Orellana et al., 2019). Für beide Zugänge gilt: Je mehr Hilfestellungen eine Person benötigt, um eine Aufgabe

bzw. einen Test erfolgreich bewältigen zu können, desto geringer ist das Lernpotenzial. Benötigt eine Person jedoch nur wenige oder sogar keine Hilfestellungen, ist das Lernpotenzial entsprechend größer (Hasson & Joffe, 2007).

Ein Kritikpunkt an diesem Zugang lautet, dass selbst Kinder mit SES in der Lage sind, unter hochstrukturierten Bedingungen (wie beim *interventionist approach*) sprachliche Aufgaben zu lösen. Es sind die natürlichen Situationen, in denen Schwierigkeiten auftreten und nicht überwunden werden können. Kinder mit SES könnten im Graduated Prompting also durchaus Lernpotenzial demonstrieren, obwohl ihnen dies unter alltäglichen Bedingungen weniger gelingt (Gutiérrez-Clellen & Peña, 2001). Sowohl das Test-Teach-Retest-Verfahren als auch das Graduated Prompting weisen Einschränkungen in ihrer diagnostischen Aussagekraft auf. Durch die Kombination beider Ansätze erhofft man sich jedoch differenziertere und zuverlässigere Ergebnisse (Hunt et al., 2022).

Einen dritten Zugang zur Messung des Lernpotenzials stellen die *Modifiability Scales* dar. Hierbei handelt es sich um Beobachtungsbögen, mit denen sich das kognitive und sozial-emotionale Verhalten des Kindes während der Testsituation einschätzen lässt (Lam et al., 2024). Petersen et al. (2017) entwickelten den *Modifiability Index,* der aus sieben Aussagen besteht, die auf einer dreistufigen Skala (niemals – manchmal – meistens) eingeschätzt werden müssen. Dabei handelt es sich um Aussagen wie „Das Kind hat auf die Prompts angesprochen", „Das Kind war aufmerksam" oder „Das Kind war nicht sichtlich frustriert" usw. Gemäß der Metaanalyse von Orellana et al. (2019) sind es v.a. Skalen wie diese, die am eindeutigsten zur Differenzierung von Kindern mit und ohne SES beitragen. Häufig werden Modifiability Scales ergänzend zu Test-Teach-Retest- bzw. Graduated Prompting-Verfahren eingesetzt und tragen so zur Erhöhung der diagnostischen Validität bei (Schwob et al., 2024).
Betrachtet man DA aus einer testpsychologischen Perspektive, ist offensichtlich, dass das Kriterium der Durchführungsobjektivität nicht gegeben ist bzw. sogar bewusst aufgegeben wird, da die Interaktion der Testleitung mit der Testperson wesentlicher Bestandteil des Vorgehens ist. Auch die Auswertungsobjektivität ist nicht zwangsläufig gegeben, welche sich durch Maße der Interrater-Reliabilität bemessen lässt. Demnach sollten zwei Personen unabhängig voneinander zu den gleichen Punkt- oder Leistungswerten kommen. Während in Tests und Fragebögen Auswertungsschablonen genutzt werden können, ist dies im Kontext von Verhaltensratings weniger gut möglich (Bühner, 2011). Je nachdem, welchen Ansatz man verfolgt, muss entsprechend auch die Interrater-Reliabilität im DA bewusst aufgegeben werden: Im *interactionist approach* steht die individuelle Interaktion mit der Testperson im Vordergrund. Wird diese korrekt umgesetzt, müssten entsprechende Tests bei zwei unterschiedlichen Personen zwangsläufig zu unterschiedlichen Verläufen und auch zu unterschiedlichen Ergebnissen führen, d.h. die Interrater-Reliabilität fällt entsprechend gering aus. Die Ergebnisse werden erst wieder vergleichbar, wenn das Vorgehen standardisiert ist, wie dies beim *interventionist approach* der Fall ist. Die Standardisierung des Vorgehens erhöht die Interrater-Reliabilität, geht allerdings zu Lasten der Individualisierung der Messung (Dumas et al., 2020).

1.3 Dynamic Assessment in der Wortschatzdiagnostik

Vor dem Hintergrund der oben skizzierten Komplexität der Wortschatzdiagnostik bei sukzessiv mehrsprachigen Kindern besteht die Erwartung, dass DA einen vielversprechenden Ansatz darstellen könnte, um die genannten Schwierigkeiten zu adressieren. Ehlert (2021) listet insgesamt 13 Studien auf, die semantisch-lexikalische Leistungen als Gegenstand von DA hatten. Die Studien von Maragkaki (2021), MacLeod und Glaspey (2022), Böse und Elstrodt-Wefing (2023) und Schwob et al. (2024) sind später erschienen und ebenfalls im Bereich Semantik-Lexik angesiedelt. Jedoch berichten die wenigsten dieser Studien Werte für Sensitivität und Spezifität, um die Differenzierungsleistungen der untersuchten Vorgehensweisen einzuschätzen (Schwob et al., 2024). Kapantzoglou et al. (2012) haben ein dynamisches fast mapping-Verfahren entwickelt und untersucht. Auch wurden Angaben zu Sensitivität und Spezifität ermittelt. Da die vorliegende Studie wie auch diejenigen von MacLeod und Glaspey (2022) und Schwob et al. (2024) sich stark am Vorgehen von Kapantzoglou et al. (2012) orientieren, soll diese Studie hier exemplarisch vorgestellt werden.
Im Test-Teach-Retest-Ansatz von Kapantzoglou et al. (2012) wurden insgesamt 28 sukzessivbilinguale Kinder im Alter zwischen vier und fünf Jahren untersucht, davon 15 ohne und 13

mit SES. In der Testphase wurden die Kinder aufgefordert, insgesamt sechs Bilder zu benennen. Drei der Bilder stellten Alltagsbegriffe dar (Blume, Pizza, Sonnenbrille), drei stellten eher alltagsfremde Objekte dar (ein unbekanntes Tier, Samenkörner und eine Wasserwaage), die die Kinder nicht benennen konnten. Diesen Objekten wurden Pseudowörter zugeordnet (fote, depa, kina), die eine vergleichbare und einfache CVCV-Struktur aufweisen. Diese Wörter wurden den Kindern in der darauffolgenden Teachphase beigebracht. Dabei handelte es sich um eine nach einem festgelegten Skript ablaufende Spielsequenz mit Handpuppen *(interventionist approach)*. Eine der Handpuppen hatte Geburtstag und die anderen Handpuppen schenkten ihr die sechs genannten Objekte. Den Kindern wurde immer wieder deutlich gemacht, um was es bei dem Spiel gehe *(„Now, we are going to play and learn some new words. Try to remember the names.")*, um die Involviertheit der Kinder in die Aufgabe zu steigern. Die Testleitung ordnete die zu merkenden Wörter dabei in verschiedene Kategorien ein, beschrieb sie, erklärte die Funktion und präsentierte dabei die Wörter mindestens neun Mal, wobei sie das Kind mindestens dreimal zum Nachsprechen des Wortes anregte. In der Retest-Phase schließlich mussten die Kinder die Wörter benennen bzw. wiedererkennen.

Das Vorgehen benötigte etwa 35 bis 40 Minuten. Zusätzlich wurden Modifiability Scales ausgewertet. Die Interrater-Reliabilität bezüglich der verschiedenen Testergebnisse betrug zwischen 91 % und 100 % Übereinstimmung, bezüglich der Modifiability Scales 94 % und 97 %. Die Auswertung ergab, dass die Ergebnisse aus dem Benenntest, dem Wiedererkennen und den Modifiability Scales zusammengenommen zu einer Sensitivität von 76.9 % und einer Spezifität von 80 % führen und damit knapp den Ansprüchen an differenzialdiagnostische Instrumente genügen. Kapantzoglou et al. (2012) bewerten dieses Ergebnis vor dem Hintergrund der Komplexität von Wortschatzdiagnostik bei sukzessiv-mehrsprachigen Kindern aber als durchaus positiv.
Schwob et al. (2024) führten ein ähnliches Verfahren durch und verglichen dieses mit einer automatisierten Computerdiagnostik, ohne jedoch eine Form der Modifiability Scales zu prüfen. Sie erreichten bezüglich der Testergebnisse eine Interrater-Reliabilität von Cohens Kappa = .98. Bei den Werten für Sensitivität und Spezifität wurden Werte von 100 % bzw. 96 % erreicht, wenn alle Ergebnisse, d.h. sowohl die des interaktiven Spiels als auch die der automatisierten Diagnostik, zur Leistungsdifferenzierung herangezogen wurden. Da es in der Praxis jedoch sehr aufwendig wäre, beide Tests durchzuführen, berechneten sie nochmals die Werte für die interaktive Spielsequenz allein. Die Sensitivität lag dann bei 86 % und die Spezifität bei 88 %, was immer noch über den Werten von Kapantzoglou et al. (2012), aber nach wie vor unter dem empfohlenen Wert von 90 % liegt.
Die Überprüfung der Wortlernleistungen von bilingualen Kindern mit Methoden, die dem DA zuzurechnen sind, scheint vielversprechend. Spielsituationen, die einem Skript folgen und somit dem interventionist approach zuzuordnen sind, erzielen hohe Werte bezüglich der Interrater-Reliabilität. Die Differenzierung von Kindern mit und ohne SES liegt derzeit jedoch nur im „zufriedenstellenden" Bereich – der Wunschwert von 90 % bezüglich Sensitivität und Spezifität wurde knapp verfehlt. Die Kombination von Test-Teach-Retest- und Graduated-Prompting-Elementen einerseits und Modifiability Scales andererseits könnte eine Verbesserung der Werte ermöglichen. Nach diesen Empfehlungen wurde das Verfahren „Nili hat Geburtstag" Anfang 2023 entwickelt und erprobt. Ob dieses Verfahren auch die Ansprüche an die Interrater-Reliabilität und die diagnostischen Differenzierungsleistungen erfüllt, ist Gegenstand der vorliegenden Untersuchung.

2 Fragestellungen

Das entwickelte DA-Verfahren des Wortlernens orientiert sich stark an Kapantzoglou et al. (2012), vereint Test-Teach-Retest- und Graduated Prompting-Elemente und folgt dabei einem Skript und ist somit dem *interventionist approach* zuzuordnen. Zusätzlich werden Modifiability Scales (genauer der Modifiability Index von Petersen et al., 2017) eingesetzt. Im Rahmen der Erprobung dieses Instruments sollen folgende Fragestellungen beantwortet werden:

Fragestellung 1: Wie hoch ist die Interrater-Reliabilität des DA-Verfahrens für sukzessiv mehrsprachige Kinder im Bereich Wortschatz bezüglich der verschiedenen Elemente (Test-Teach-Retest mit aktiver und passiver Wortschatzüberprüfung, Graduated Prompting, Modifiability Index)?

Fragstellung 2: Wie gut differenzieren das Gesamtverfahren und seine verschiedenen Elemente zwischen mehrsprachigen Kindern mit logopädischem und ohne logopädischen Förderbedarf?

a) Unterscheiden sich Kinder mit logopädischem und ohne logopädischen Förderbedarf im Gesamtverfahren und in den Elementen signifikant voneinander?
b) Wie hoch sind Sensitivität und Spezifität des Gesamtverfahrens und seiner einzelnen Elemente?

3 Methode

3.1 Untersuchungsdesign

Die Durchführung der Studie erfolgte in Zusammenarbeit mit insgesamt 20 Logopädinnen des logopädischen Dienstes der Stadt St. Gallen. Bis Ende März 2023 führten diese das Verfahren „Nili hat Geburtstag" mit jeweils zwei Kindern im ersten oder zweiten Kindergartenjahr durch. Alle Kinder sprechen zu Hause eine nicht-deutsche Erstsprache und Deutsch als Zweitsprache in der Umgebung.

Nachdem das Einverständnis der Eltern eingeholt wurde, wurde die Durchführung des DA-Verfahrens videografiert. Unmittelbar darauf, spätestens jedoch nach 14 Tagen, wurde außerdem der Subtest 1 „Bildbenennung" des SET 5-10 (Petermann, 2018) durchgeführt. Alle Videos der Durchführung des DA-Verfahrens wurden von zwei Personen unabhängig voneinander mittels eines Beurteilungsprotokolls bewertet.

3.2 Stichprobe

Die Stichprobe bestand aus insgesamt 40 mehrsprachig aufwachsenden Kindern, die im Schnitt zwischen 5;7 und 6;2 Jahren alt waren. Die Daten eines Kindes konnten jedoch aufgrund technischer Fehler bei der Videoaufnahme nicht ausgewertet werden (s. Tab. 1). Fast alle Kinder in der Stichprobe besuchen das zweite Kindergartenjahr. Der Kindergarten ist in der Schweiz Teil der obligatorischen Schulzeit und umfasst 20-24 Wochenlektionen. Dies stellt die Mindestmenge an Deutschkontakt dar, welchen die Kinder pro Woche bekommen. Zum Zeitpunkt der Untersuchung betrug das Minimum an Deutschkontakt also sechs Monate, in der Regel aber mindestens 18 Monate. Von den 39 Kindern befinden sich 28 Jungen und Mädchen in logopädischer Behandlung, 11 Kinder werden nicht logopädisch betreut. Zu den Kindern, die sich in logopädischer Behandlung befinden, gaben die Logopädinnen die Hauptergebnisse der logopädischen Abklärung (Testresultate und Diagnosen) an. Aus diesen Angaben wurde ersichtlich, dass sich drei Kinder aufgrund einer rein phonetischen Aussprachestörung in logopädischer Therapie befanden. Diese drei Kinder wurden nachträglich der Gruppe der nicht logopädisch behandelten Kinder zugeordnet, da hier nicht von einer Sprachentwicklungsproblematik ausgegangen werden kann. Auf die Verwendung des Terminus Sprachentwicklungsstörung (SES) wird im Folgenden bewusst verzichtet, da die Diagnosen in der Praxis nicht auf standardisierten und wissenschaftlich abgesicherten Verfahren beruhen. Stattdessen wird zwischen Kindern mit und ohne logopädischen Förderbedarf differenziert. Die Gruppe der Kinder mit logopädischem Förderbedarf umfasst somit 25 Kinder, die Gruppe der Kinder ohne logopädischen Förderbedarf 14 Kinder. Die Kinder mit logopädischem Förderbedarf sind im Durchschnitt 5 Monate älter als die Kinder ohne logopädischen Förderbedarf ($t(37) = 2.21$, $p < .05$), weisen aber im Wortschatztest des SET-5-10 signifikant schlechtere Ergebnisse auf ($t(36) = -2.70$, $p < .05$).

Tab. 1: Angaben zur Stichprobe

Gruppe	Anzahl	Alter (M)	Schulstufe	Geschlecht (m : w)	Wortschatzleistung (Rohwert)
Kinder in logopädischer Behandlung	25	6;2 Jahre	1. KG: 2 2. KG: 21 (Fehlend: 2)	18:7	11.20 (6.25)
Kinder ohne logopädische Behandlung	14	5;7 Jahre	1. KG: 3 2. KG: 11	8:6	17.92 (9.04)

3.3 Dynamic Assessment-Rollenspiel

In Anlehnung an Kapantzoglou et al. (2012) wurde ein Rollenspiel konzipiert („Nili hat Geburtstag“), in dessen Verlauf die Kinder sich sechs Zutaten zu einem Kuchen merken müssen (s. Abb. 1).

Abb. 1: Rezept für einen Nili-Kuchen (KI-generierte Bilder)

Drei von diesen Zutaten bzw. Wörtern sind Alltagsbegriffe (Bananen, Eier, Zucker), drei sind sog. quasi-universelle Nichtwörter (lafi, paklu, pifakup), die dem LITMUS-NWR in der Version von Grimm und Hübner (in Druck) entnommen worden sind. Diese Nichtwörter bestehen nur aus Phonemen bzw. Silben, die in den meisten Sprachen der Welt vorkommen und somit auch von den meisten Menschen dieser Welt nachgesprochen werden können. Die drei Nichtwörter wurden aufgrund ihrer vergleichbaren Länge und linguistischen Komplexität zu den Realwörtern ausgewählt und sind außerdem unterschiedlich genug, damit sie nicht miteinander verwechselt werden können. Auch bei Macleod und Glaspey (2022) und Schwob et al. (2024) wurden Nichtwörter aus dem LITMUS-NWR verwendet und auch dort wurden nur drei unbekannte Wörter erlernt, um die Kapazitäten der Kinder nicht zu überschreiten.

In einem ersten Schritt wird den Kindern gesagt, dass ein Geburtstagskuchen gebacken werden soll und dass man zu diesem Zweck verschiedene Zutaten braucht. Es wird betont, dass man sich die Namen der verschiedenen Zutaten gut merken muss und dass man dem Kind dabei helfen wird (in Anlehnung an Ehlert, 2021). Die sechs Wörter werden bildlich dargestellt, wobei die Nichtwörter durch reale Abbildungen von eher alltagsfernen Lebensmitteln repräsentiert werden (namentlich eine Kiwano, Pintobohnen und Matchapulver).

Im Prätest werden die sechs Abbildungen dem Kind in Form eines Rezepts vorgelegt, damit es diese erstmalig benennen kann. Für gewöhnlich werden die drei Alltagsbegriffe erfolgreich benannt, die Nichtwörter hingegen nicht.

Sodann werden die sechs Begriffe dem Kind erstmalig vorgesprochen (die Testleitung zeigt nacheinander auf die Bilder und sagt: „Das sind Bananen, das sind /lafi/, das sind Eier, das sind /paklu/…“ usw.). Direkt im Anschluss wird das Kind gebeten, die sechs Abbildungen mit den entsprechenden Wörtern zu bezeichnen. Es wird deutlich, ob es dem Kind gelingt, die Wörter bereits nach einmaligem Hören wiederzugeben.

Die nächsten Schritte im Skript erfolgen wortweise, d.h. zunächst mit „lafi“, dann mit „paklu“ und anschließend mit „pifakup“, sind aber immer gleich durchzuführen. So werden die Wörter zunächst semantisch elaboriert. Die semantischen Hinweise der Nichtwörter beziehen sich auf

die Realwörter, um an bereits vorhandenes Wortwissen anknüpfen zu können (z.B. "Eine lafi ist eine Frucht, so wie die Banane. Die ist aber nicht süß, sondern sauer! Auch wenn die Schale stachelig aussieht, kann man die mitessen" usw.). Als nächster, stärkerer Hinweis erfolgt die phonologische Elaboration: die Wörter werden silbisch zergliedert ("la - fi"). Das Kind wird aufgefordert, die Silben gemeinsam mit der erwachsenen Person zu klatschen, was bedeutet, dass es spätestens jetzt zum ersten Mal das Wort selbst ausspricht. Im Anschluss an die phonologische Elaboration erfolgt der Einsatz des "Zaubertricks" aus dem Wortschatzsammler (Motsch et al., 2018). Die Wörter werden dreimal langsam und deutlich ausgesprochen. Der Zaubertrick soll den Kindern dabei helfen, auch neue und schwierige Wortformen besser einspeichern zu können. Nach Einsatz des Zaubertricks hat das Kind das Nichtwort mindestens viermal selbst ausgesprochen. Diese Schritte werden nun auch für die anderen beiden Nichtwörter durchgespielt. Nach jedem dieser Schritte muss das geübte Wort eingefordert werden ("Weißt Du jetzt, wie es heißt?"), um eine Rückmeldung zu bekommen, welche Hilfestellungen dem Kind geholfen haben und welche nicht. Dies ist auch eine Gelegenheit, falsch abgespeicherte Wortformen (z.B. „paku" statt „paklu") unmittelbar zu korrigieren.
Um die Wörter vor dem Posttest noch einmal ins Bewusstsein zu rufen, wird nun das Tempospiel (Motsch et al., 2018) gespielt: innerhalb einer Minute müssen die neuen Begriffe so oft wie möglich benannt werden. Auf diese Weise sollen Abrufroutinen etabliert werden, um so die Wahrscheinlichkeit eines erfolgreichen Wortabrufs zu erhöhen. Die erwachsene Person kann und muss das Kind beim Tempospiel unterstützen, indem sie die Wortformen bei Bedarf vorsagt, damit dieses Spiel erfolgreich abgeschlossen werden kann.

Darauf folgt der Posttest: Um die Kuchenzutaten zu erhalten, muss das Kind den Lieferdienst anrufen. Die testleitende Person spielt die Verkäuferin oder den Verkäufer am anderen Ende der Leitung. Das Kind bestellt nun nach und nach sowohl die bereits bekannten als auch die neu hinzugelernten Zutaten. Das Nachspielen einer Telefonsituation zur Erzeugung eines obligatorischen, nicht durch Zeigen lösbaren Kontextes orientiert sich am Vorgehen des ESGRAF 4–8 (Motsch & Rietz, 2016). Wenn das Kind es nicht schafft, ein Wort aus dem Gedächtnis abzurufen, werden hierarchisch gestaffelte Hinweisreize gegeben, welche sich an dem Vorgehen aus dem Wortschatz- und Wortfindungstest für 6- bis 10-Jährige (Glück, 2011) orientiert. Der erste Hinweis besteht in einer Ermunterung und einem zeitlichen Aufschub ("Denk noch einmal in Ruhe nach, dir fällt es bestimmt wieder ein!"), was die Motivation des Kindes, das Wort doch noch zu finden, erhöhen soll. Wenn dies nicht gelingt, erfolgt der zweite Hinweis, bei dem die semantischen Hinweise, die in der Trainingsphase erarbeitet worden sind, nun als Unterstützung vorgegeben werden ("Das ist auch eine Frucht, so wie die Banane. Die schmeckt aber sauer. Und auch wenn die Schale stachelig aussieht, kann man die mitessen" usw.). Wenn auch dieser Hinweis nicht zum erfolgreichen Abruf führt, wird nun der Anlaut vorgegeben, um mit dem phonologischen Hinweis die stärkste Hilfestellung anzubieten. Gelingt auch dies nicht, wird das Wort vorgegeben, um den Fortschritt des Spiels zu ermöglichen ("Meinst Du vielleicht /lafi/...?").
Nachdem die Lebensmittel erfolgreich bestellt worden sind, werden diese dem Kind in Form von Bildkärtchen gegeben. Zu den sechs Zielwörtern werden zwei weitere Bildkarten (Fotos von Mandeln und Cherimoyas) als Ablenker hinzugefügt. Dem Kind wird jetzt das Rezept vorgelesen, wobei es die entsprechenden Bilder passend zu den Instruktionen in eine Rührschüssel legen soll (z.B. "Zuerst brauchen wir die Bananen. Jetzt brauchen wir /lafi/" usw.). Auf diese Weise soll in letzter Instanz das Wortverständnis überprüft werden, da die Wortproduktion für gewisse Kinder bis zuletzt eine Herausforderung bleiben dürfte. Dabei ist die Reihenfolge der Wörter gemäß den Instruktionen einzuhalten, damit immer eine Auswahl zwischen mehreren Pseudowörtern gegeben ist und das Kind nicht per Ausschlussverfahren das richtige Bild findet.
Unter Einhaltung dieses Vorgehens wird das produktive und rezeptive Lernen von neuen Wörtern überprüft, man erhält aber auch Hinweise zur Anzahl und der Art der vom Kind benötigten Hilfestellungen, die während des Worttrainings und beim Posttest für den erfolgreichen Abruf benötigt wurden. Dabei hört das Kind die Zielwörter mindestens sieben Mal, spricht sie selbst mindestens vier Mal und hat insgesamt sechs Gelegenheiten, die Wörter auf Nachfrage abzurufen. V.a. beim Tempospiel dürfte die Menge der gehörten und erfolgreich abgerufenen Zielwörter jedoch deutlich gesteigert werden.

3.4 Messinstrumente

3.4.1 Dynamic Assessment Verfahren

Die Leistung der Kinder wird innerhalb des DA-Verfahrens anhand dreier Skalen und zusätzlich anhand des Modifiability Index ermittelt. Tabelle 2 stellt die Schritte des Spielskripts und die daraus resultierenden Beurteilungsskalen im Überblick dar. Die Skalen werden nur mit den Pseudowörtern gebildet, da die drei realen Wörter in der Regel auch Kindern mit sehr geringen Deutschkenntnissen bekannt sind. Die Skala zum ersten Teil des Spiels kann unter dem Begriff des Graduated Promptings zusammengefasst werden, da hier Hinweisreize zum Wortlernen in einer vorgegebenen Hierarchie präsentiert werden und nach jedem Schritt die Abrufbarkeit der neuen Wörter geprüft wird. Die aktive Wortschatzprüfung im Telefonspiel entspricht prinzipiell dem „Retest" im Test-Teach-Retest-Paradigma, allerdings werden auch hier Elemente des Graduated Promptings integriert.

Tab. 2: Durchführung des Verfahrens und Beurteilungsskalen im Überblick

Was	Wie	Beurteilungsskalen	Anzahl der Items und der erreichbaren Punkte
Einführung der neuen Wörter	Das Nili-Rezept wird gemeinsam angeschaut, bekannte Wörter werden vom Kind, unbekannte Wörter von der Logopädin benannt.	Graduated Prompting: alle Punkte, die im Verlaufe der Wortlernphase erreicht wurden	12 Items, maximal 24 Punkte
Überprüfung 1	Welche Items können spontan wieder reproduziert werden?		
Elaboration der neuen Wörter	Semantische und phonologische Elaboration a) Einführung semantischer Merkmale; b) Silbensegmentation; c) Wörter 3x Nachsprechen		
Überprüfung 2a)-c)	Welche Items können nach jeder Elaboration unmittelbar benannt werden?		
Tempospiel	1 Minute Benennen der neuen Wörter.		
Überprüfung 3	Die Zutaten müssen telefonisch im Supermarkt bestellt werden. Die Logopädin gibt bei Bedarf Hinweise in vorgegebener Reihenfolge (allgemein, semantisch, phonologisch).	Retest - Aktive Wortschatzüberprüfung, inklusive Elemente des Graduated Prompting	3 Items, maximal 12 Punkte. Für jedes Item wird bewertet: 0 = selbst mit Prompt nicht abrufbar 1 = mit Hilfe angenähert 2 = mit Hilfe richtig 3 = ohne Hilfe angenähert 4 = ohne Hilfe richtig
Überprüfung 4	Die Logopädin gibt vor, in welcher Reihenfolge die Zutaten in die Schüssel gegeben werden müssen. Es liegen zusätzlich unbekannte Bilder auf dem Tisch (Ablenker).	Retest - Rezeptive Wortschatzüberprüfung	3 Items, maximal 6 Punkte
Beobachtung	Der Modifiability Index wird nach der Spielsequenz ausgefüllt..	7 Verhaltensaspekte werden auf einer vierstufigen Skala bewertet.	7 Items, maximal 28 Punkte
Gesamt		Gesamtpunktwert aus den drei Skalen des DA-Verfahrens + Modifiability Index	maximal 70 Punkte

In beiden Skalen (Graduated Prompting, aktive Wortschatzüberprüfung) wird erfasst, ob das Kind die richtige Wortform abrufen konnte oder zumindest eine ähnliche Wortform. Für diese Unterscheidung werden die Anweisungen von Grimm und Hübner (im Druck) zum LITMUS-NWR umgesetzt. Das Zielwort gilt als korrekt, wenn es vollständig oder mit geringfügigen Abweichungen (Plosive stimmhaft statt stimmlos und umgekehrt) wiedergegeben wurde. Annähernd korrekt ist die Wortform, wenn die Abfolge der Vokale oder der Konsonanten erhalten geblieben ist oder auch wenn eine Silbe verwechselt oder ersetzt, nicht aber, wenn sie ausgelassen wurde. Sind die Abweichungen größer, gilt die Wortform als inkorrekt. Bei der aktiven Wortschatzüberprüfung wird zusätzlich notiert, ob der Abruf direkt oder erst nach einer der Hil-

festellungen (motivierender, semantischer oder phonologischer Hinweis) gelang. Diese beiden Aspekte wurden in das in Tabelle 2 ersichtliche Punktesystem überführt.
In der dritten Skala, dem rezeptiven Wortschatztest werden zwei Punkte für die korrekte Zuordnung erteilt und ein Punkt, wenn ein Kind sich für ein anderes Pseudowort entscheidet, da man so erkennen kann, dass das Kind die richtige Lösung bei den Nichtwörtern, und nicht bei den Realwörtern sucht (Ausschlussprinzip).
Insgesamt kann das Kind eine Maximalpunktzahl von 42 erreichen. 18 Punkte sind allein in Prä- und Posttest enthalten. Die restlichen 24 Punkte können während des Worttrainings erworben werden. Sie geben gleichzeitig darüber Aufschluss, welche und wie viele Hilfestellungen für das Kind benötigt wurden, um einen erfolgreichen Wortabruf zu ermöglichen (Glück, 2011). Die Auswertung des Verfahrens erfolgt anhand eines vorgegebenen Beurteilungsprotokolls.

3.4.2 Modifiability Index

Neben den drei Skalen des DA-Verfahrens selbst wird ein Beobachtungsbogen in Anlehnung an den Modifiability Index von Petersen et al. (2017) eingesetzt. Dieser wurde zunächst ins Deutsche übersetzt und teilweise angepasst, um die Beschreibungen treffender auf die oben beschriebene Wortlernsituation zu beziehen. Außerdem wurde eine vierte Skalenstufe ergänzt, da in der Pilotierungsphase bei nur drei Skalenstufen eine deutliche Tendenz zur Mitte beobachtet werden konnte. Insgesamt werden sieben Verhaltensaspekte bewertet (Reaktionen auf Hilfestellungen, Ausmaß der Transferleistungen, Aufmerksamkeit bei Lernaufgabe, Lernprozess, Frustration, Störungen und Gesamteindruck). Bei Petersen et al. (2017) stehen niedrige Werte (Minimum 1) für positive Verhaltensweisen und hohe Werte (Maximum 4) für negative Verhaltensweisen. In der vorgenommenen Untersuchung wurde die Skala umgepolt, damit hohe Werte (wie im Rest des Verfahrens) für ein positives Ergebnis stehen und niedrige Werte für ein negatives Ergebnis. Somit sind sieben Punkte der Minimalwert und 28 Punkte der Maximalwert, der erreicht werden kann.
Die Skalenenden werden durch beispielhafte Beschreibungen illustriert. Das Beispiel für positive Verhaltensaspekte auf der Skala für "Störungen" lautet: "Das Kind zeigt kein ablenkendes oder störendes Verhalten (verbal und/oder nonverbal), ist kooperativ, ist nicht motorisch unruhig, …". Das Negativbeispiel lautet hingegen: "Das Kind zeigt ablenkendes oder störendes Verhalten (verbal und/oder nonverbal), ist unkooperativ (Verweigerung, Vermeidung), ist während der Testung motorisch unruhig (zappelt, dreht sich weg, steht auf) …".
Alle Materialien, das verwendete Skript und das Beurteilungsprotokoll sind im Download-Bereich des Artikels verfügbar.

3.4.3 Wortschatztest aus dem SET 5-10

Zusätzlich zum DA-Verfahren wurde der Bildbenenntest aus dem SET 5-10 (Petermann, 2018) zu Beginn der Untersuchung durchgeführt. Insgesamt handelt es sich um 40 Begriffe, davon 28 Nomen und 12 Verben. Die Kinder erhalten einen Punkt, wenn der Begriff korrekt oder mit einem gemäß Protokoll gültigen Synonym bezeichnet wurde. Null Punkte gibt es für falsche Begriffe, Bezeichnungen von Teilen der Abbildung, für Umschreibungen, Eigenschöpfungen sowie unspezifische Universalwörter ("Ding"). Die Itemschwierigkeiten liegen zwischen 0.14 (sehr schwer) und 0.98 (sehr leicht). Fünf- bis sechsjährige Kinder erzielen im Mittel zwischen 24.31 bzw. 32.40 Punkten. Die in der Stichprobe ermittelten Werte der Kinder mit DaZ liegen deutlich darunter. Dieser Befund ist nicht unerwartet, da mehrsprachig aufwachsende Kinder in ihren einzelnen Sprachen oft einen kleineren Wortschatz aufweisen als ihre einsprachig aufwachsenden Peers, was sich aber relativiert, wenn man alle ihre Sprachen berücksichtigt (Byers-Heinlein et al., 2024; Hoff et al., 2012; Montanari et al., 2018).

3.5 Datenauswertung

Bevor die Daten zwecks Beantwortung der Forschungsfragen analysiert werden konnten, mussten die fehlenden Daten (15 % Missings) imputiert werden. Der Hauptgrund für fehlende Daten war, dass die durchführenden Logopädinnen das Benennen der Items im Teach-Teil des Verfahrens nicht konsequent einforderten und somit keine Bewertung vorgenommen werden konnte. Es wurden 15 Imputationen durchgeführt, weil dies dem durchschnittlichen Prozentsatz der fehlenden Werte entspricht (White et al., 2011). Die 15 imputierten Datensätze wurden daraufhin in einen gepoolten Datensatz überführt, der die Grundlage für alle folgenden Berechnungen darstellt.

Für die Berechnung der Interrater-Reliabilität wird Cohen's Kappa (κ) angegeben, welches für zwei Rater und kategoriale Daten empfohlen wird (Hallgreen, 2012). Zusätzlich wird auch Krippendorff's Alpha (α) berichtet, da dieses im Umgang mit fehlenden Daten und kleinen Stichproben als besonders geeignet gilt (Krippendorff, 2004). Die Zusammenschau beider Werte ermöglicht somit eine differenzierte Einschätzung. Für die Interpretation gilt jeweils, dass Werte unter .67 als nicht zuverlässig gelten, Werte zwischen .67 und .80 nur vorläufige Schlüsse erlauben, und erst ab .80 verlässliche Aussagen möglich sind (Krippendorff, 2004).

Die Gruppenvergleiche zur Beantwortung von Fragestellung 2a) erfolgen mittels t-Tests für unabhängige Stichproben. Da eine klare Hypothese zur erwarteten Richtung der Resultate formuliert werden kann (Kinder mit logopädischem Förderbedarf schneiden schlechter ab als Kinder ohne logopädischen Förderbedarf), wird einseitig getestet.

Sensitivität und Spezifität (Fragestellung 2b) werden mittels ROC-Analysen (ROC: Receiver Operating Characteristic) bestimmt, um die diagnostische Genauigkeit des Verfahrens zur Unterscheidung zwischen Kindern mit einem und ohne logopädischen Förderbedarf zu ermitteln. Dabei bezeichnet die Sensitivität den Anteil der Kinder mit logopädischem Behandlungsbedarf, die durch das Verfahren korrekt identifiziert werden, während die Spezifität angibt, wie viele Kinder ohne Behandlungsbedarf korrekt als unauffällig klassifiziert werden. Die Area Under the Curve (AUC) dient als zusammenfassendes Maß für die Trennschärfe eines Verfahrens. Diese Auswertungen erfolgen nur für Skalen, in denen die zuvor durchgeführten t-Tests (Fragestellung 2a) signifikante Gruppenunterschiede dokumentieren. Sensitivitäts- und Spezifitätswerte von ≥ 90 % gelten als gut, Werte zwischen 80–89 % als angemessen und Werte unter 80 % als unzureichend (Shahmahmood et al., 2016). Für die AUC gelten Werte von ≥ 0.90 als exzellent, 0.80–0.89 als gut, 0.70–0.79 als akzeptabel, 0.60–0.69 als schwach und < 0.60 als unzureichend (Çorbacıoğlu und Aksel, 2023).

4 Ergebnisse

4.1 Interrater-Reliabilität

Die 39 Videos des DA-Verfahrens wurden von zwei Personen unabhängig voneinander ausgewertet. Die Übereinstimmung zwischen den Urteilen der beiden Personen in den einzelnen Subskalen ist in Tabelle 3 dokumentiert. Die Reliabilitätsanalyse zeigt eine unterschiedliche Übereinstimmung zwischen den Personen in Abhängigkeit der verwendeten Skala. Für jede Skala wurde die Interrater-Reliabilität (Krippendorff's α und Cohen's κ) auf Item-Ebene berechnet. Die im Folgenden berichteten Werte geben jeweils den Mittelwert dieser itembezogenen Reliabilitätskennwerte pro Skala an. Zusätzlich werden die Minimal- und Maximalwerte zur Darstellung der Streuung innerhalb der Skalen ausgewiesen.

Die höchste Übereinstimmung wurde bei der rezeptiven Wortschatzüberprüfung erzielt. Hier lagen die mittleren Werte für Krippendorff's α (M = .84) und Cohen's κ (M = .82) im Bereich über .80, mit relativ enger Streuung (α: .73–.95; κ: .79–.84). Auch die aktive Wortschatzüberprüfung zeigte insgesamt hohe Mittelwerte (α: .82; κ: .72), allerdings mit einer breiteren Spannweite (α: .70–.98; κ: .56–.96).

Bei der Skala Graduated Prompting lagen die Mittelwerte im mittleren Bereich (α: M = .76; κ: M = .72), wobei die Streuung deutlich größer war (α: .57–.97; κ: .54–.92).

Der Modifiability Index zeigten die geringste Interrater-Reliabilität. Die itembasierten Mittelwerte lagen für Krippendorff's α bei .50 und für Cohen's κ bei .40. Die Spannbreite reichte von .29 bis .79 (α) bzw. von .16 bis .68 (κ), was auf erhebliche Unterschiede in der Beurteilbarkeit einzelner Items innerhalb dieser Skala hinweist. Die höchste Übereinstimmung innerhalb des Modifiability Index erzielte die Gesamteinschätzung des kindlichen Verhaltens.

Tab. 3: Interrater-Reliabilität der Skalen des DA-Verfahrens

Skala	Anzahl Items	Krippendorffs Alpha M (Min – Max)	Cohens Kappa M (Min – Max)
Graduated Prompting	12	.76 (.57 - .97)	.72 (.54 - .92)
Aktive Wortschatzüberprüfung	3	.82 (.70 - .98)	.72 (.56 - .96)
Rezeptive Wortschatzüberprüfung	3	.84 (.73 - .95)	.82 (.79 - .84)
Modifiability Index	7	.50 (.29 - .79)	.40 (.16 - .68)*

* Beim Modifiability Index wurde das gewichtete lineare Kappa berechnet

Die Analyse zeigt außerdem, dass niedrige Werte der Interrater-Reliabilität beim Graduated Prompting insbesondere bei jenen Items auftreten, bei denen Uneinigkeit darüber bestand, ob das Verhalten vom Kind tatsächlich evoziert worden war. In diesen Fällen lag häufig eine Bewertungslücke bei Person 2 vor – d. h., eine Person vergab Punkte, während die andere das Item nicht berücksichtigte. Da meist angenommen werden konnte, dass das Verhalten gezeigt wurde, wenn eine Person eine Bewertung vergab, wurde für die weiteren Analysen (Fragestellungen 2a und 2b) in solchen Fällen die Einschätzung der bewertenden Person (in der Regel Person 1) übernommen.

4.2 Gruppenunterschiede zwischen Kindern mit und ohne logopädischen Förderbedarf

Zur Überprüfung möglicher Leistungsunterschiede zwischen Kindern mit logopädischem Förderbedarf und ohne logopädischen Förderbedarf wurden unabhängige t-Tests für die verschiedenen Skalen durchgeführt (s. Tab. 4). Die Ergebnisse zeigen signifikante Gruppenunterschiede in mehreren Bereichen zugunsten der Kinder ohne logopädischen Förderbedarf.
Im Gesamtscore, der alle dynamischen Testverfahren umfasst, erzielte die Gruppe ohne Logopädiebedarf signifikant höhere Werte (p = .014), mit einer mittleren Effektstärke (d = –.76). Auch im Modifiability Index zeigten sich signifikante Unterschiede zugunsten der Kinder ohne Logopädiebedarf (p = .011, d = –.79). Die aktive Wortschatzüberprüfung, die gemeinsam mit der rezeptiven Überprüfung die Skala Retest bildet, ergab ebenfalls einen signifikanten Gruppenunterschied (p = .039, d = –.60). Für die rezeptive Wortschatzüberprüfung sowie für den zusammengesetzten Retest-Score ergaben sich keine signifikanten Unterschiede (p > .05). Ein ähnliches Bild zeigt sich für die Skala DA, die sich aus der Wortlernphase (Graduated Prompting) und dem Retest zusammensetzt: Für beide Komponenten sowie für die zusammengesetzte Skala ergaben sich keine signifikanten Unterschiede zwischen den Gruppen (alle p > .05).

Tab. 4: Mittelwertvergleiche zwischen den Gruppen mit/ohne Logopädie in den Skalen des DA-Verfahrens

	Mit Logopädie (M, SD)	Ohne Logopädie (M, SD)	t (37), p, d		Mit Logopädie	Ohne Logopädie	t (37), p, d		Mit Logopädie	Ohne Logopädie	t (37), p, d		Mit Logopädie	Ohne Logopädie	t (37), p, d
Wortlernphase (Graduated Prompting)	13.71 (4.00)	15.53 (4.32)	t = -1.32 p = .097					Skala Dynamic Assessment				Gesamtscore			
Aktive Wortschatz-überprüfung	4.15 (3.07)	6.17 (3.78)	t = -1.81 p = .039 d = -.60	Skala Retest	9.24 (3.63)	11.42 (4.38)	t = -1.67 p = .051		23.04 (6.26)	26.78 (7.39)	t = -1.68 p = .051		42.47 (8.98)	49.64 (10.20)	t = -2.28 p = .014 d = -.76
Rezeptive Wortschatz-überprüfung	5.08 (1.14)	5.28 (0.91)	t = -.56 p = .289												
Modifiability Index	19.52 (3.84)	22.64 (4.05)	t = -2.39 p = .011 d = -.79												

Anmerkung: p-Werte basieren auf einseitigen Tests; farblich hinterlegte Zellen enthalten signifikante Ergebnisse

4.3 Sensitivität und Spezifität des Verfahrens

ROC-Analysen wurden für jene Skalen des Verfahrens durchgeführt, bei denen in den vorgängigen Gruppenvergleichen (t-Tests) signifikante Unterschiede zwischen Kindern mit logopädischem und ohne logopädischen Förderbedarf festgestellt wurden. Dies betraf die Skalen Aktive Wortschatzüberprüfung, Modifiability Index sowie den Gesamtwert (s. Tab. 5).
Die aktive Wortschatzüberprüfung zeigte mit einer AUC von .654 eine nur geringe diagnostische Genauigkeit, die zudem statistisch nicht signifikant war (p = .103). Damit ist die Trennschärfe dieser Skala im Hinblick auf den Förderbedarf eingeschränkt. Der optimale Cut-Off ergab eine moderate Sensitivität und Spezifität (jeweils ca. 64 %).
Im Gegensatz dazu erreichte der Modifiability Index eine deutlich bessere Trennschärfe (AUC = 0.723; p = .013). Diese Skala konnte Kinder mit und ohne Förderbedarf mit guter Genauigkeit unterscheiden. Der entsprechende Schwellenwert bot ein ausgewogenes Verhältnis zwischen Sensitivität (71,4 %) und Spezifität (67,9 %).

Auch der Gesamtwert erwies sich mit einer AUC von .703 als signifikanter Prädiktor für logopädischen Förderbedarf (p = .022). Bemerkenswert ist hier insbesondere die hohe Spezifität von 96,0 % beim optimalen Cut-Off, während die Sensitivität mit 42,9 % vergleichsweise niedrig ausfiel. Dies spricht dafür, dass der Gesamtwert besonders gut geeignet ist, Kinder ohne Förderbedarf korrekt zu identifizieren. Er ist jedoch weniger sensitiv gegenüber einem vorliegenden Förderbedarf.

Tab. 5: Diagnostische Kennwerte (ROC) der drei analysierten Skalen

Skala	AUC	95 %-KI	p-Wert	Cut-Off	Sensitivität (%)	Spezifität (%)	Youden-Index
Aktive Wortschatz-überprüfung	.654	.469–.840	.103	5.33	64.3	64.0	.283
Modifiability Index	.723	.546–.899	.013	21.5	71.4	67.9	.394
Gesamtwert	.703	.529–.877	.022	54.87	42.9	96.0	.389

Anmerkung: AUC = Area Under the Curve

5 Diskussion

Das hier präsentierte DA-Verfahren der Wortschatzdiagnostik für sukzessiv mehrsprachige Kindergartenkinder integriert verschiedene Elemente: Es beinhaltet sowohl Graduated Prompting als auch ein Test-Teach-Retest-Design und die sich der Testsituation anschließende Einschätzung des kindlichen Verhaltens durch den Modifiability Index. Aufgrund der engen Vorgaben zu Art, Anzahl und Reihenfolge der Hinweisreize ist das Verfahren dem *interventionist approach* zuzuordnen. Überprüft wurde zum einen, ob die Standardisierung des Vorgehens eine Vergleichbarkeit der Ergebnisse zwischen verschiedenen beurteilenden Personen ermöglicht und zum anderen, ob das Verfahren zwischen Kindern mit logopädischem und ohne logopädischen Förderbedarf zu differenzieren vermag.

Die oben präsentierten Ergebnisse bieten einen differenzierten Blick auf die Interrater-Reliabilität der verschiedenen Skalen. Die Skalen des DA-Verfahrens selbst (Graduated Prompting, aktive und rezeptive Wortschatzüberprüfung) überschreiten die gewünschte Mindestanforderung von .80 (Krippendorff, 2004) entweder oder liegen knapp unter diesem Wert. V.a. die Werte bezüglich des Graduated Prompting unterliegen großen Schwankungen, da zum Teil Unklarheit darüber bestand, ob das Zielwort vom Kind aktiv abgerufen oder nur nach- bzw. mitgesprochen wurde und ob also Punkte gegeben oder nicht gegeben werden konnten. Trotz ausführlichem Skript unterschied sich die Testdurchführung bei den teilnehmenden Logopädinnen teilweise erheblich. Somit scheint – auch in einem interactionist approach – die Durchführungsobjektivität nicht zwangsläufig gegeben zu sein.

Die Werte für die Interrater-Reliabilität des Modifiability Index liegen mit .50 (Krippendorff‘s α) bzw. .40 (Cohen‘s κ) sehr deutlich unter den gewünschten Mindestanforderungen. Im Modifiability Index (angelehnt an Petersen et al., 2017) besteht jede Kategorie aus einem Oberbegriff und aus Beobachtungshinweisen, die global eingeschätzt werden. Wie in Abschnitt 3.3.2 bereits gezeigt, werden dabei unterschiedliche Verhaltensweisen in einem Item zusammengefasst, und ein einzelner Verhaltensaspekt kann nicht eindeutig beurteilt werden bzw. verschiedene Personen bewerten unterschiedliche Verhaltensaspekte aus derselben Kategorie mit anderer Gewichtung. Einzig bei der Beurteilung des Gesamteindrucks *(Wie schätzen Sie – basierend auf Ihrer Interaktion mit dem Kind - die Wortlernfähigkeit des Kindes ein?)* wurden gute Übereinstimmungswerte (Krippendorf’s α = .79) erreicht. Den eingangs erwähnten Ausführungen von Bühner (2011) folgend, ist das Erreichen einer hohen Auswertungsobjektivität und damit einer zufriedenstellenden Interrater-Reliabilität besonders im Kontext von Verhaltensratings weniger gut möglich. Verhalten wird häufig in Form von sog. Verhaltensankern festgehalten, deren Auftreten bewertet wird. „Fehler treten auf, wenn die Auswerter nicht das gleiche Verständnis der Verhaltensanker haben“ (Bühner, 2011, S. 59). Zukünftige Anwendungen des Modifiability Index sollten entsprechend das zu beobachtende Verhalten genauer operationalisieren und verschiedene Aspekte deutlich voneinander abgrenzen. Vor dem Hintergrund der eigenen Befunde bezüglich der nied-

rigen Interrater-Reliabilität ist zudem in künftigen Studien eine gründliche Vorbereitung auf die Anwendung entsprechender Beobachtungsinstrumente empfehlenswert (Lam et al., 2024). Lam et al. (2024) weisen zusätzlich darauf hin, dass die Struktur des Modifiability Index von Petersen et al. (2017) nie wissenschaftlich evaluiert worden ist. Somit bleibt unklar, ob mittels dieses Instruments relevante Verhaltensweisen bzw. Verhaltensanker, die einen Aufschluss über das Sprachlernpotenzial geben können, beurteilt werden. Einzig der Aufbau der Mediated Learning Observation von Peña et al. (2007) wurde mittels konfirmatorischer Faktoranalysen von Lam et al. (2024) überprüft, generell bestätigt und leicht angepasst. In künftigen Studien wird daher empfohlen auf dieses angepasste Instrument zurückzugreifen.

Auch für die zweite Fragestellung ist es lohnend, die verschiedenen Subskalen des Verfahrens einzeln zu betrachten. Sukzessiv mehrsprachig aufwachsende Kindergartenkinder mit logopädischem und ohne logopädischen Förderbedarf unterscheiden sich in der aktiven Wortschatzüberprüfung (Telefonspiel mit Graduated Prompting), im Modifiability Index und im Gesamtscore (alle Punkte im DA + Modifiability Index) signifikant voneinander, nicht jedoch in der Wortlernphase, in der rezeptiven Wortschatzüberprüfung und in der Gesamtpunktzahl des DA-Verfahrens ohne Modifiability Index. Der Modifiability Index, der auch in den Gesamtwert mit einfließt, scheint die Abgrenzung zwischen beiden Gruppen hingegen besonders zu begünstigen. Dies verweist auf die Bedeutung der genauen Beobachtung des kindlichen Verhaltens während der Interaktion, zunächst einmal unabhängig vom erzielten Punktwert.

Die nicht-signifikanten Resultate im Untertest Graduated Prompting (Wortlernphase) und in der rezeptiven Wortschatzüberprüfung legen nahe, dass auch Kinder mit logopädischem Förderbedarf in hochstrukturierten Wortlernsituationen gute Ergebnisse erzielen können und dass auch das passive Wiedererkennen der eingeführten Wörter – zumindest kurzfristig - für diese Kinder möglich ist, wie bereits Gutiérrez-Clellen und Peña (2001) anmerken. Erst bei steigenden Anforderungen wie z.B. beim eigenständigen freien Abruf der neuen Items während der aktiven Wortschatzüberprüfung (Telefonspiel) werden signifikante Gruppenunterschiede offensichtlich. Die Ergebnisse der ROC-Analysen für die drei untersuchten Skalen (Aktive Wortschatzüberprüfung, Modifiability Index, Gesamtwert) bewegen sich zwischen 43 % und 71 % und sind damit als unzureichend einzustufen (Shahmahmood et al., 2016). Ausnahme ist die Spezifität im Gesamtwert (96 %): Kinder ohne Förderbedarf werden somit zuverlässig in die richtige Gruppe eingeteilt. Die Sensitivität liegt bei dieser Skala allerdings bei 43 %, so dass Kinder mit Förderbedarf mit hoher Wahrscheinlichkeit falsch klassifiziert werden.

Zudem muss eine wichtige Grundsatzfrage diskutiert werden. Die Theorie besagt, dass mit dynamischen Verfahren das (Sprach-)Lernpotenzial und nicht das (Sprach-)Wissen überprüft wird (Hasson & Joffe, 2007). Allerdings zeigt sich, dass es v.a. die Kinder sind, die gemäß dem Wortschatztest des SET 5-10 bereits über ein genügend großes semantisch-lexikalisches Wissen verfügen, die auch im dynamischen Verfahren besonders erfolgreich sind. Die Kinder, die niedrige Ergebnisse im SET 5-10 erzielen und über eher wenig semantisch-lexikalisches Wissen verfügen, sind auch im dynamischen Verfahren weniger erfolgreich. Mit r = .448 korrelieren die Rohwerte des SET 5-10 mittelstark mit dem Gesamtergebnis des DA-Verfahrens, mit r = .581 liegt sogar eine hohe Korrelation mit dem Ergebnis des Modifiability Index vor. Mit einem AUC = .719 liegt die diagnostische Trennschärfe des Bildbenenntests in einem ähnlich hohen Bereich wie die des Modifiability Index.
Die oben geschilderte Beobachtung stellt die Möglichkeit in Frage, das Sprachlernpotenzial von Kindern, die sukzessiv mehrsprachig aufwachsen, unabhängig vom vorhandenen Sprachwissen zu erheben. Der eingangs beschriebene Matthäus-Effekt (je größer das vorhandene Wortwissen, desto schneller werden neue Wörter hinzugelernt; Juska-Bacher et al., 2023; Rupp, 2021) spielt bei der vorliegenden Untersuchung eventuell eine entscheidende Rolle. Künftige Untersuchungen sollten den Zusammenhang zwischen bereits vorhandenem Sprachwissen und Sprachlernpotenzial gezielt untersuchen.
Für den Einsatz in der Praxis kann das hier dokumentierte Verfahren aufgrund der aufgeführten Kritikpunkte in der aktuellen Form nicht empfohlen werden. Eine Weiterentwicklung ist aber in verschiedener Hinsicht möglich: Neben dem Einsatz der angepassten Mediated Learning Observation (Lam et al. 2024) statt des hier verwendeten Modifiability Index liegt Potenzial in den folgenden beiden Modifikationen:

1) Die Wortlernsequenz mittels Graduated Prompting vermochte nicht zwischen Kindern mit und ohne Förderbedarf zu differenzieren, vermutlich weil auch Kinder mit SES in hochstrukturierten Lernsequenzen erfolgreich sein können. Bei diesem Teil des Skripts könnten den durchführenden Logopädinnen folglich mehr Freiheiten eingeräumt werden. Bei gleichbleibender Art der Hilfestellungen (semantische Elaboration, phonologische Elaboration, Zaubertrick) könnte dann jede Logopädin eine Spielsequenz kreieren, die den Fähigkeiten des Kindes und ihrem eigenen Stil angepasst ist.
2) Bei der Auswertung der Videos wurde lediglich die erfolgreiche Wortproduktion dokumentiert. Möglicherweise spiegelt sich das Sprachlernpotenzial eines Kindes aber weniger im Punktestand, sondern vielmehr in seinem Verhalten während der Sprachlehr-/-lernsituation wider. Dieses Verhalten wird in allgemeiner Art und Weise durch Modifiability Scales bewertet. Allerdings könnte auch die Beobachtung von Verhaltensweisen bzw. Verhaltensankern interessant sein, die nachweislich den Worterwerbsprozess begünstigen, z.B. das Nachsprechen von vorgegebenen Wörtern oder die selbständige Verknüpfung mit eigenem Vorwissen.

In der Studie von Böse und Elstrodt-Wefing (2023) wurden differenzierte Kriterien zum Verhalten eines Kindes in der Wortlernsituation entwickelt und erprobt. Wird die Graduated-Prompting-Phase gemäß Vorschlag 1) angepasst, ergeben sich hier viele Gelegenheiten, das spezifische Wortlernverhalten einzuschätzen.

6 Limitationen

Zu den Schwächen der hier dargestellten Untersuchung gehört sicherlich die eher kleine Stichprobe, die sich in zwei ungleich große Gruppen unterteilt. Ein verfehltes Signifikanzniveau könnte im vorliegenden Fall also auch durch die kleine Stichprobengröße und somit durch reduzierte Testpower bedingt sein, zumal statistische Signifikanz sowohl in der Retest-Skala als auch in der Gesamtskala des DA-Verfahrens jeweils nur knapp verpasst wurde (p=.051).
Zudem besteht auch eine signifikante Altersdifferenz zwischen den beiden Gruppen. Diese wurde im Kontext der statistischen Analysen jedoch bewusst nicht berücksichtigt, da sich der Altersunterschied vor dem Hintergrund der SET-Ergebnisse nicht als Vorteil für die älteren Kinder erwiesen hat. Dennoch besteht die Möglichkeit, dass die älteren Kinder über mehr kognitive Strategien verfügen, die sich im Erlernen der neuen Wörter günstig auswirken.
Auch die Einteilung in die Gruppen „mit Förderbedarf“ und „ohne Förderbedarf“ muss hinterfragt werden. Diese wurde aufgrund der Beurteilungen der zuständigen Logopädinnen vorgenommen – eine Überprüfung dieser Einteilung mittels standardisierter Verfahren der Sprach- und/oder Intelligenzdiagnostik konnte nicht vorgenommen werden. Die Interpretation der Ergebnisse muss auch vor dem Hintergrund der fehlenden differenzierten Informationen über Länge und Qualität des Kontaktes mit der deutschen Sprache noch einmal relativiert werden.
Im geplanten Studiendesign war außerdem vorgesehen, den Wortschatztest des SET 5-10 nach einem sechsmonatigen Intervall zu wiederholen, um den Lernfortschritt der Kinder mit dem Abschneiden im DA-Verfahren in Beziehung zu setzen. Da diese Daten aber nur für 25 von 39 Teilnehmende vorliegen, können entsprechende Auswertungen dazu nicht verlässlich vorgenommen werden. Somit bleibt unklar, ob durch die Zuteilung in die Gruppen (logopädischer Förderbedarf ja / nein) tatsächlich die Kinder identifiziert wurden, die trotz ausreichender Inputgelegenheiten in ihrer Wortschatzentwicklung (eben wegen eines mangelnden Sprachlernpotenzials) zurückbleiben.

7 Implikationen für die Praxis

Die Eignung des Verfahrens in seiner jetzigen Form zur Differenzierung der Kinder mit und ohne Förderbedarf wird durch die ROC-Analysen nicht unterstützt, die Gefahr für Fehldiagnosen ist weiterhin groß. Trotz aller gefundenen Nachteile des Verfahrens „Nili hat Geburtstag“ ist aber auch das Potenzial des Vorgehens deutlich geworden. Nicht zuletzt sind Belege aus dem Bereich der sozialen Evidenz zu nennen. So hatten die beobachteten Kinder sichtlich Spaß am Spiel und wirkten nicht, als ob sie getestet wurden. Und auch die Logopädinnen schilderten, dass sie die ihnen bekannten Kinder noch einmal neu kennenlernen und neue Seiten an ihnen entdecken konnten. Sie durften die Rolle der objektiven Testleitung, die das Manual streng befolgt, verlas-

sen und die Kinder während der Testung unterstützen, was ihrer therapeutischen Grundhaltung mehr entsprach. Dies ist zwar auch in standardisierter Diagnostik möglich, allerdings dürfen die Antworten dann nicht mehr gewertet werden. Im DA wird dieses Vorgehen jedoch legitimiert, bewusst genutzt und bekommt somit einen besonderen Stellenwert. Gelingt es, diese praktischen Vorteile zu nutzen und die hier aufgedeckten Nachteile auszugleichen, liegt ein brauchbares Diagnostikinstrument vor, dessen Potenzial künftig auch praktisch genutzt werden könnte.

Literaturverzeichnis

Bonuck, K., Shafer, V., Battino, R., Valicenti-McDermott, R. M., Sussman, E. S. & McGrath, K. (2022). Language Disorders Research on Bilingualism, School-Age, and Related Difficulties: A Scoping Review of Descriptive Studies. *Academic pediatrics, 22*(4), 518–525. doi: 10.1016/j.acap.2021.12.002

Böse, J. & Elstrodt-Wefing, N. (2023). Dynamic Assessment zur Erfassung von Fast-Mapping-Prozessen und Spracherwerbsstrategien bei drei- bis fünfjährigen Kindern im anfänglichen Deutschspracherwerb. *Praxis Sprache, (1)*, 16-23.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson.

Byers-Heinlein, K., Gonzalez-Barrero, A. M., Schott, E. & Killam, H. (2024). Sometimes larger, sometimes smaller: Measuring vocabulary in monolingual and bilingual infants and toddlers. *First Language, 44*(1), 74–95. doi: 10.1177/01427237231204167

Çorbacıoğlu, Ş. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine, 23*(4), 195-198. https://doi.org/10.4103/tjem.tjem_182_23

Dixon, C., Hessel, A., Smith, N., Nielsen, D., Wesierska, M. & Oxley, E. (2022). Receptive and expressive vocabulary development in children learning English as an additional language: Converging evidence from multiple datasets. *Journal of child language, 1–22*. doi: 10.1017/S0305000922000071

Dumas, D., McNeish, D. & Greene, J. A. (2020). Dynamic measurement: A theoretical–psychometric paradigm for modern educational psychology. *Educational Psychologist, 55*(2), 88–105. doi: 10.1080/00461520.2020.1744150

Ehl, B., Schrey-Dern, D. & Willmes, K. (2014). Der AWST-R bei sukzessiv mehrsprachigen Kindern: Eignung und Anpassung der Auswertung bei sukzessiven Erwerbsbedingungen. *Forum Logopädie, 28*(1), 30–34.

Ehlert, H. (2021). *Dynamic Assessment. Prozess und Potential in der Diagnostik von Sprachentwicklungsstörungen.* Wiesbaden: Springer.

Freeman, M. R. & Schroeder, S. R. (2022). Assessing Language Skills in Bilingual Children: Current Trends in Research and Practice. *Journal of Child Science, 12*(1), e33-e46. doi: 10.1055/s-0042-1743575

Glück, C.W. (2011). *Wortschatz- und Wortfindungstest für 6 bis 10-Jährige (WWT 6-10)* (2. Aufl.). München: Urban & Fischer.

Grimm, A. & Hübner, J. (in Druck). *Nonword repetition by bilingual learners of German: The role of language-specific complexity.*

Gutiérrez-Clellen, V. F. & Peña, E. D. (2001). Dynamic Assessment of Diverse Children: A Tutorial. *Language, Speech and Hearing Services in Schools, (32)*, 212–224.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34. doi: 10.20982/tqmp.08.1.p023

Hasson, N. & Joffe, V. (2007). The case for Dynamic Assessment in speech and language therapy. *Child Language Teaching and Therapy, 23*(1), 9–25.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M. & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language, 39*(1), 1–27. doi: 10.1017/S0305000910000759

Holzinger, D., Weber, C. & Jezek, M. (2022). Identifying Language Disorder Within a Migration Context: Development and Performance of a Pre-school Screening Tool for Children With German as a Second Language. *Frontiers in pediatrics, 10,* 1–11. doi: 10.3389/fped.2022.814415

Hunt, E., Nang, C., Meldrum, S. & Armstrong, E. (2022). Can Dynamic Assessment Identify Language Disorder in Multilingual Children? Clinical Applications From a Systematic Review. *Language, Speech, and Hearing Services in Schools, 53*(2), 598–625. doi: 10.1044/2021_LSHSS-21-00094

Juska-Bacher, B., Brugger, L. & Lingg, M. (2024). *Wie Drittklässler:innen beim Lesen unbekannte Wörter entschlüsseln oder „einfach schnell geraten"?* Bielefeld: wbv Media.

Kapantzoglou, M., Restrepo, M. A. & Thompson, M. S. (2012). Dynamic Assessment of Word Learning Skills: Identifying Language Impairment in Bilingual Children. *Language, Speech, and Hearing Services in Schools, 43*(1), 81–96. doi: 10.1044/0161-1461(2011/10-0095)

Klassert, A. & Kauschke, C. (2014). Semantisch-lexikalische Entwicklungsstörungen bei mehrsprachigen Kindern. In S. Chilla & S. Haberzettl (Hrsg.), *Handbuch Spracherwerb und Sprachentwicklungsstörungen* (S. 121–134). Stuttgart: Elsevier.

Kauschke, C., Lüke, C., Dohmen, A., Haid, A., Leitinger, C., Männel, C., Penz, T., Sachse, S., Scharff Rethfeldt, W., Spranger, J., Vogt, S., Neumann, K. & Niederberger, M. (2023). Delphi-Studie zur Definition und Terminologie von Sprachentwicklungsstörungen - eine interdisziplinäre Neubestimmung für den deutschsprachigen Raum. *Logos, 31*(1), 2–20.

Krippendorff, K. (2004). Reliability in Content Analysis. *Human Communication Research, 30*(3), 411–433. doi: 10.1111/j.1468-2958.2004.tb00738.x

Lam, J. H. Y., Resendiz, M. D., Bedore, L. M., Gillam, R. B. & Peña, E. D. (2024). Validation of the Mediated Learning Observation Instrument Among Children With and Without Developmental Language Disorder in Dynamic Assessment. *Journal of Speech, Language, and Hearing Research : JSLHR, 67*(7), 2159–2171. doi: 10.1044/2024_JSLHR-23-00127

Lautenschläger, T., Sawatzky, A., Schneller, K., Kaiser-Kratzmann, J., Kierdorf, J. & Sachse, S. (2023). Sprachentwicklungsverläufe bei mehrsprachigen Kindern im Vorschulalter: Durchschnittliche Leistungsveränderungen und interindividuelle Unterschiede in der Entwicklung der Umgebungssprache zwischen dem Alter von drei und sechs Jahren. *Forschung Sprache, (2)*, 86–107.

MacLeod, A. A. N. & Glaspey, A. M. (2022). Dynamic assessment of multilingual children's word learning. International *Journal of Language & Communication Disorders, 57*(4), 822–851. doi: 10.1111/1460-6984.12723

Maragkaki, I. (2021). *Dynamic Assessment of receptive vocabulary and phonology of preschool children with German as a second language.* Université de Genève. doi: 10.13097/ARCHIVE-OUVERTE/UNIGE:154220

Montanari, E. G., Abel, R., Graßer, B. & Tschudinovski, L. (2018). Do bilinguals create two different sets of vocabulary for two domains? *Linguistic Approaches to Bilingualism, 8*(4), 502–522. doi: 10.1075/lab.16021.mon
Motsch, H.-J., Marks, D.-K. & Ulrich, T. (2018). *Wortschatzsammler: Evidenzbasierte Strategietherapie lexikalischer Störungen im Kindesalter* (3. Aufl.). Sprachtherapie. München: Ernst Reinhardt.
Motsch, H.-J., & Rietz, C. (2016). *ESGRAF 4–8: Grammatiktest für 4- bis 8-jährige Kinder – Manual* (1. Auflage). München: Ernst Reinhardt.
Orellana, C. I., Wada, R. & Gillam, R. B. (2019). The Use of Dynamic Assessment for the Diagnosis of Language Disorders in Bilingual Children: A Meta-Analysis. *American Journal of Speech-Language Pathology, 28*(3), 1298–1317. doi: 10.1044/2019_AJSLP-18-0202
Peña, E. D., Reséndiz, M. & Gillam, R. B. (2007). The role of clinical judgements of modifiability in the diagnosis of language impairment. *International Journal of Speech-Language Pathology, 9*(4), 332–345. doi: 10.1080/14417040701413738
Petermann, F. (2018). *SET 5-10: Sprachstandserhebungstest für Kinder im Alter zwischen 5 und 10 Jahren* (3. Aufl.). Göttingen: Hogrefe.
Petersen, D. B., Chanthongthip, H., Ukrainetz, T. A., Spencer, T. D. & Steeve, R. W. (2017). Dynamic Assessment of Narratives: Efficient, Accurate Identification of Language Impairment in Bilingual Students. *Journal of Speech, Language, and Hearing Research : JSLHR, 60*(4), 983–998. doi: 10.1044/2016_JSLHR-L-15-0426
Rothweiler, M. & Ruberg, T. (2011). *Der Erwerb des Deutschen bei Kindern mit nichtdeutscher Erstsprache: Sprachliche und außersprachliche Einflussfaktoren.* München: Dt. Jugendinst.
Rupp, S. (2021). *Wortschatzförderung bei früh sukzessiv deutschlernenden Kindern: Eine modellgeleitete empirische Untersuchung.* Düren: Shaker.
Scherger, A.-L. (2023). Diagnostik mehrsprachiger Kinder in der sprachtherapeutischen Praxis – Stand der Forschung in Über- und Ausblick. *Sprachtherapie aktuell: Forschung – Wissen – Transfer,* 1–13.
Schwob, S., Tillé, Y. & Skoruppa, K. (2024). A comparison of two dynamic assessment situations for detecting development language disorder in monolingual and bilingual children. *Clinical Linguistics & Phonetics,* 1–19. doi: 10.1080/02699206.2024.2435010
Shahmahmood, T. M., Jalaie, S., Soleymani, Z., Haresabadi, F. & Nemati, P. (2016). A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues. *Journal of Research in Medical Sciences : the Official Journal of Isfahan University of Medical Sciences, 21*, 67. doi: 10.4103/1735-1995.189648
Sheng, L., McGregor, K. K. & Marian, V. (2006). Lexical–Semantic Organization in Bilingual Children: Evidence From a Repeated Word Association Task. *Journal of Speech Language and Hearing Research, 49*(3), 572. doi: 10.1044/1092-4388(2006/041)
Thordardottir, E. T. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism, 15*(4), 426–445. doi: 10.1177/1367006911403202
White, I. R., Royston, P. & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377–399. doi: 10.1002/sim.4067
Wilkens, R., Lein, T. & Rothweiler, M. (2018). Sprachdiagnostik bei zweisprachigen Kindern: Phonologische Verarbeitung und Wortschatzleistungen. *Praxis Sprache, 63*(1), 24–30.

Danksagung

Wir danken Dr. Nadine Itel, der Leiterin des logopädischen Dienstes der Stadt St. Gallen, sowie ihrem Team für die tatkräftige Unterstützung. Natürlich sei auch all den Kindern gedankt, die uns bei „Nili hat Geburtstag“ geholfen haben.

Korrespondenzadressen

Dr. Christoph Till
PHBern
Fabrikstrasse 8, CH-3012 Bern
christoph.till@phbern.ch

Dr. Julia Winkes
Departement für Sonderpädagogik der Universität Freiburg (CH)
Petrus-Kanisius-Gasse 21, CH-1700 Freiburg
julia.winkes@unifr.ch

Informationen zu den Autor:innen

Christoph Till ist Bereichsleiter für die Fachwissenschaften und Dozent für Sprachheilpädagogik am Institut für Heilpädagogik der PHBern. Seine Forschungsinteressen liegen im Spracherwerb mehrsprachiger Kinder sowie bei der multiprofessionellen Zusammenarbeit von Regellehrpersonen, schulischen Heilpädagog:innen und Logopäd:innen.

Julia Winkes ist Lektorin am Departement für Sonderpädagogik der Universität Freiburg (CH) in den Abteilungen Logopädie und Schulische Heilpädagogik. Sie lehrt und forscht zum Sprach- und Schriftspracherwerb im Schulalter und zur Lernverlaufsdiagnostik.

