

Erfassung von mathematischen Kompetenzen im Vorschulalter mit MARKO-D

Erprobung des Tests MARKO-D an einer Schweizer Stichprobe und Überprüfung des Instruments für Messwiederholungen

Ariana Garrote¹  und Elisabeth Moser Opitz²

¹Pädagogische Hochschule, Fachhochschule Nordwestschweiz

²Institut für Erziehungswissenschaft, Universität Zürich, Schweiz

Zusammenfassung: In dieser Studie wurde der Test MARKO-D (Mathematik- und Rechenkonzepte im Vorschulalter–Diagnose) mit einer Stichprobe von Kindern aus der deutschsprachigen Schweiz ($N = 555$) im ersten und zweiten Kindergartenjahr erprobt und es wurde analysiert, ob sich die Altersnormen der deutschen Stichprobe auf die Schweiz übertragen lassen. Zudem wurde der Test mit einer Teilstichprobe ($n = 87$) hinsichtlich Messinvarianz über die Zeit untersucht. Die Ergebnisse des eindimensionalen Rasch-Modells zeigen, dass das Instrument für die Schweiz geeignet ist. Die Testleistungen hängen jedoch vom Kindergartenbesuch ab. Für die Schweiz müssten deshalb neben Altersnormen auch Normen pro Kindergartenhalbjahr verwendet werden. Die Analyse mittels Differential Item Functioning ergab, dass 17 von 55 Items von großer Messvarianz über die Zeit betroffen sind. Um das Instrument für Längsschnittuntersuchungen einsetzen zu können, müsste es weiterentwickelt werden.

Schlüsselwörter: Mathematische Kompetenzen, Kindergarten, Normierung, Messwiederholung

Assessing Mathematical Competence of Preschoolers With MARKO-D: Validating and Testing the Instrument's Suitability for Longitudinal Assessment in Switzerland

Abstract: In this study, the test MARKO-D (Math and Calculation – A Test for Diagnosing Concepts at Pre-school Age) was validated in a sample of preschoolers enrolled in the first and second year of kindergarten in Switzerland ($N = 555$) and the applicability of the age norms from Germany for a Swiss sample was analyzed. In addition, measurement invariance over time was analyzed in a subsample ($n = 87$). A unidimensional Rasch model revealed that the instrument is suitable for Switzerland. However, mathematical competence also depends on the time spent in kindergarten. Therefore, norms should be based on age as well as on kindergarten experience. Differential item functioning analysis revealed that 17 out of 55 items were affected by measurement variance over time. The instrument needs to be adapted for use in longitudinal studies.

Keywords: mathematical competence, kindergarten, standardisation, longitudinal assessment

Erfassen von mathematischen Kompetenzen im Vorschulalter

Die mathematische Förderung von Kindern im Vorschulalter ist zu einem wichtigen Desiderat geworden. Dazu beigetragen haben Studien, die zum einen die große Bedeutung der numerischen Kenntnisse für die spätere mathematische Entwicklung und zum anderen die Förderbar-

keit dieser Kompetenzen aufgezeigt haben (Gallit et al., 2018; Hauser, Vogt, Stebler & Rechsteiner, 2014; Krajewski & Schneider, 2009; Langhorst, Hildenbrand, Ehlert, Ricken & Fritz, 2013; Missall, Mercer, Martinez & Casebeer, 2012). In diesem Kontext kommt den Instrumenten, mit denen im Kindergarten die mathematische Entwicklung untersucht werden kann, eine hohe Bedeutung zu. Allerdings entstehen bei der Leistungsmessung im Kindergarten Herausforderungen auf unterschiedlichen Ebe-

nen. Die Leistungen der Kinder sind sehr heterogen und hängen sowohl mit individuellen Merkmalen wie Intelligenz (Hauser et al., 2014) oder Geschlecht (Anders, 2012; Kuratli Geeler, 2019) als auch mit kontextuellen Faktoren wie dem sozioökonomischen Status (Schuchardt, Piekny, Grube & Mähler, 2014), der Erstsprache (Anders, 2012) oder dem Kontext der vorschulischen Förderung (Fritz, Ehlert & Leutner, 2018; Kuratli Geeler, 2019) zusammen. In der vorliegenden Studie interessiert der Zusammenhang mit dem Kindergartenbesuch in der Schweiz.

Im letzten Jahrzehnt sind mehrere deutschsprachige Tests bzw. Testreihen erschienen, mit denen mathematische Kompetenzen im Kindergarten bzw. der Schuleingangsphase erfasst werden können (Kaufmann et al., 2009; Krajewski, 2018; Ricken, Fritz & Balzer, 2013). Es gibt jedoch bisher keine Instrumente mit Normen für die Schweiz und es stellt sich erstens die Frage, ob sich die Normen aus Deutschland auf die Schweiz übertragen lassen. Dies ist deshalb bedeutsam, weil sich die pädagogischen Konzepte des Kindergartens in Deutschland und der Schweiz unterscheiden (Gasteiger, Brunner & Chen, 2020). In der Schweiz gehört der Kindergarten zur Volksschule, ist unentgeltlich und – mit Ausnahme eines Kantons – mindestens für ein Jahr obligatorisch. Fast alle Kinder besuchen den Kindergarten für zwei Jahre. Im Lehrplan 21 (Deutschschweizer Erziehungsdirektorenkonferenz [D-EDK], 2014) sind verbindliche fachliche Lernziele für das mathematische Lernen formuliert. In Deutschland ist der Kindergarten freiwillig, die Eltern bezahlen in den meisten Fällen Beiträge und viele Trägerschaften sind privat (Gasteiger et al., 2020). Die Bildungs- und Orientierungspläne unterscheiden sich inhaltlich stark von Bundesland zu Bundesland und in einigen Fällen wird die mathematische Förderung nur am Rande erwähnt (Gasteiger et al., 2020).

Damit verbunden ist zweitens nicht geklärt, ob für den Vorschulbereich Altersnormen geeignet sind, oder ob der Kindergartenbesuch und damit verbunden die potentielle mathematische Förderung, die in den normalen Kindergartenalltag integriert wird, ebenfalls berücksichtigt werden müsste. Die Wirksamkeit von spezifischer, sowohl trainings- wie auch spielbasierter Förderung wurde in mehreren Studien nachgewiesen (für den deutschsprachigen Raum z.B. Fritz, Ehlert & Leutner, 2018; Krajewski & Schneider, 2009; Jörns, Schuchardt, Grube & Mähler, 2014; Vogt, Hauser, Stebler, Rechsteiner & Urech, 2018). Die Frage ist jedoch, ob sich die mathematische Förderung, wie sie normalerweise gemäß dem Lehrplan im Kindergarten (in der Schweiz) stattfindet – d.h. ohne kontrollierte Trainingsstudie – auf die mathematische Entwicklung der Kinder auswirkt. Wenn das der Fall ist, müsste der Kindergartenbesuch bei der Normierung berücksichtigt werden. Dies ist für die Schweiz an-

gesichts der starken Bildungsorientierung und der Tatsache, dass fast alle Kinder den Kindergarten für zwei Jahre besuchen, besonders bedeutsam. In den vorhandenen Instrumenten wird die Normierung unterschiedlich vorgenommen. Im Test zur Erfassung numerisch-rechnerischer Fertigkeiten vom Kindergarten bis zur 3. Klasse (TEDI-MATH; Kaufmann et al., 2009) sind die Normen auf die Kindergartenhalbjahre bezogen. Im Test Mathematik- und Rechenkonzepte im Vorschulalter – Diagnose (MARKO-D; Ricken et al., 2013) stehen Altersnormen zur Verfügung mit der zusätzlichen Möglichkeit, eine entwicklungs-niveaubezogene Auswertung vorzunehmen. Für den Test mathematischer Basiskompetenzen im Kindergartenalter (MBK 0; Krajewski, 2018) gibt es sowohl Altersnormen als auch Normen für das erste, zweite und dritte Drittel des letzten Kindergartenjahres.

Drittens erfolgte die Normierung der vorhandenen Tests für den Kindergarten – abgesehen von der Re-Test-Reliabilität – oft querschnittlich an verschiedenen Stichproben. Zum Teil wurden die Instrumente längsschnittlich eingesetzt, aber es fehlen Angaben zur Messinvarianz über die Zeit (Hauser et al., 2014; Jörns et al., 2014; Krajewski & Schneider, 2009; Lyons, Bugden, Zheng, de Jesus & Ansari, 2018). Wenn dasselbe Instrument mehrmals eingesetzt werden soll, muss dieses Items enthalten, die ein breites Leistungsspektrum abdecken. Das ist im Vorschulalter besonders wichtig, weil in dieser Altersphase der Entwicklungsstand sehr heterogen und das Entwicklungstempo sehr hoch ist. Aunio, Heiskari, van Luit und Vuorio (2015) berichten, dass Skalen zum Verständnis von Mengen und Relationen sowie zum Zahlwissen und zum Zählen, die an drei Messzeitpunkten im Kindergarten eingesetzt wurden, einen Deckeneffekt aufwiesen. Zudem ist für Längsschnittuntersuchungen wichtig, dass die zeitliche Messinvarianz überprüft wird. Ist ein Instrument auf den ersten Messzeitpunkt (z.B. Anfang zweites Kindergartenjahr) hin ausgerichtet, ist es bei einer zweiten Messung (z.B. Ende zweites Kindergartenjahr) mit hoher Wahrscheinlichkeit zu einfach. Umgekehrt führt eine Anpassung an den zweiten Messzeitpunkt wahrscheinlich dazu, dass die Aufgaben bei der ersten Erfassung zu schwierig sind. Weiter muss berücksichtigt werden, dass sich Messfehler bei wiederholten Messungen und der Berechnung von Differenzwerten kumulieren, was zu einer sinkenden Reliabilität führt (Rost, 2004). Hinzu kommt, dass die Messung des interessierenden Konstrukts unabhängig vom Faktor Zeit erfolgen muss. Das heißt, dass auf der Indikatorebene des Instruments keine Interaktion mit der Variable Zeit vorliegen darf (Putnick & Bornstein, 2016).

Insgesamt lassen sich somit drei Forschungslücken identifizieren. Erstens fehlen Informationen darüber, ob die in Deutschland standardisierte Instrumente und ihre Normen für die Erfassung der mathematischen Kompe-

tenzen von Vorschulkindern aus der Schweiz eingesetzt werden können. Zweitens wurde bisher nicht untersucht, ob für den Einsatz von Tests im Kindergarten Normen notwendig sind, mit denen auch der Einfluss des Kindergartenbesuchs berücksichtigt wird. Drittens fehlen für den Vorschulbereich Instrumente zur längsschnittlichen Erhebung mathematischer Kompetenzen, bei denen die Messinvarianz über die Zeit überprüft worden ist. Diese Forschungsfragen werden in der vorliegenden Studie am Beispiel des Tests MARKO-D (Ricken et al., 2013) untersucht. Das Instrument ist dafür sehr geeignet, da es auf einem theoretisch und empirisch abgestützten Entwicklungsmodell beruht (Fritz, Ehlert & Balzer, 2013; Fritz et al., 2018; Ricken et al., 2013) und zudem eines der wenigen Instrumente ist, das auf der Basis der Item-Response-Theorie mit dem dichotomen Rasch-Modell analysiert wurden. Im Folgenden werden Erkenntnisse zur Entwicklung arithmetischer Kompetenzen, auf denen der Test MARKO-D beruht, dargestellt.

Entwicklung arithmetischer Kompetenzen

Zur Entwicklung mathematischer Kompetenzen liegt eine Vielzahl von empirischen Studien vor. Als zentrale Prädiktoren für die arithmetische Entwicklung gelten spezifisch numerische Kompetenzen, die ab dem zweiten Lebensjahr erworben werden: Zählkompetenzen (Aunio & Räsänen, 2016; Stock, Desoete & Roeyers, 2007), der symbolische Mengenvergleich (Lyons et al. 2018), die Identifikation arabischer Zahlen (Göbel, Watson, Lervåg & Hulme, 2014) sowie die Verbindung von Zahlen und Anzahlen (Krajewski & Schneider, 2009). Ausgehend von diesen Untersuchungen wurden kognitiv-entwicklungspsychologische Modelle zur Entwicklung arithmetischer Kompetenzen entwickelt (Fritz et al., 2018). Im deutschsprachigen Raum sind dies insbesondere das Modell der Zahlen-Größen-Verknüpfung von Krajewski (Krajewski & Schneider, 2009; Schneider, Küspert & Krajewski, 2016) sowie das Entwicklungsmodell arithmetischer Kompetenzen aus den Arbeitsgruppen von Ricken und Fritz (Fritz et al., 2013; Fritz et al., 2018). Der hier interessierende Test MARKO-D basiert auf dem letztgenannten Modell, das im Folgenden dargestellt wird.

Das Modell beschreibt die Entwicklung numerischer Kompetenzen von Kindern im Alter von vier bis acht Jahren anhand von sechs Niveaus (Fritz et al., 2013; Fritz et al., 2018; Ricken et al., 2013). Auf Niveau I (Zählzahl) werden Zahlen nach ihrer Mächtigkeit unterschieden. Die Kinder erwerben Zählkompetenzen und kleine Mengen können erfasst werden. Auf Niveau II wird eine unpräzise Repräsentation des Zahlenstrahls erworben und Kinder können Vorgänger- und Nachfolgerzahlen bestimm-

men. Niveau III umfasst das Verständnis der Kardinalität und der Zerlegbarkeit. Zählzahl und Anzahl werden miteinander verbunden. Zahlen und Mengen können ohne das Herstellen von Eins-zu-Eins-Zuordnung miteinander verglichen werden. Dieses kardinale Verständnis ist mit ersten Einsichten in das Zerlegen und Zusammensetzen von Mengen verbunden. Die Ausdifferenzierung dieser Einsicht steht im Zentrum von Niveau IV und kann von konkreten Mengen auf Zahlen übertragen werden (z. B. $5 = 3+2$). Niveau V umfasst das Verstehen von Relationalität und das Verbinden von Ordinalität (Niveau II), Kardinalität (Niveau III) und Relationalität (Niveau IV). Das beinhaltet das Erkennen der Zahlwortreihe als eine Sequenz aufeinanderfolgender Zahlen und das Verstehen, dass jede Zahl zwei „Nachbarzahlen“ hat, eine die um eins größer und eine, die um eins kleiner ist. Auf Niveau VI erfolgt die Einsicht, dass Zahlen gleichmächtig gebündelt werden können. Im Test MARKO-D wurden zu den Niveaus I bis V Testaufgaben entwickelt, die im Kapitel *Methodik* dargestellt werden. Bezogen auf die Schweiz, stellt sich die Frage, ob diese Entwicklungsniveaus auch in einer Schweizer Stichprobe wiederzufinden sind.

Forschungsziel

Ausgehend von den beschriebenen Forschungslücken verfolgt die hier präsentierte Studie drei Zielsetzungen: Erstens wird am Beispiel des in Deutschland entwickelten Tests MARKO-D (Ricken et al., 2013) untersucht, ob der Test bzw. die einzelnen Items in einer Schweizer Stichprobe dieselbe/n Schwierigkeit/en aufweisen wie in der Normierungsstichprobe und ob die Einteilung in die fünf Entwicklungsniveaus übernommen werden kann. Zweitens interessiert vor dem Hintergrund der unterschiedlichen pädagogischen Ausrichtungen des Kindergartens in Deutschland und in der Schweiz, ob sich die für eine deutsche Stichprobe geltenden Altersnormen auf die Schweiz übertragen lassen oder ob allenfalls zusätzlich Normen, welche die Kindergartenenerfahrung mitberücksichtigen, notwendig sind. Drittens wird mittels zeitlicher Messinvarianz überprüft, ob sich der Test für längsschnittliche Messungen über ein Kindergartenjahr hinweg und damit für die Erfassung von Lernfortschritten im Vorschulalter eignet.

Methodik

Design

Die Erhebung der Daten mit dem Test MARKO-D wurde an einer Gelegenheitsstichprobe von Kindergartenkindern

aus der Schweiz an zwei Messzeitpunkten durchgeführt (siehe Abbildung 1). Beim ersten Messzeitpunkt (t_0), am Anfang des Kindergartenjahres, nahmen Kinder im ersten und zweiten Kindergartenjahr teil. Für die Erhebung am zweiten Messzeitpunkt (t_1) ca. sieben Monate später wurden nur Kinder am Ende des zweiten Kindergartenjahres getestet, da insbesondere interessierte, ob der Test kurz vor Schuleintritt noch genügend differenziert bzw. ob Messinvarianz hinsichtlich dieses Messzeitpunkts gegeben ist.

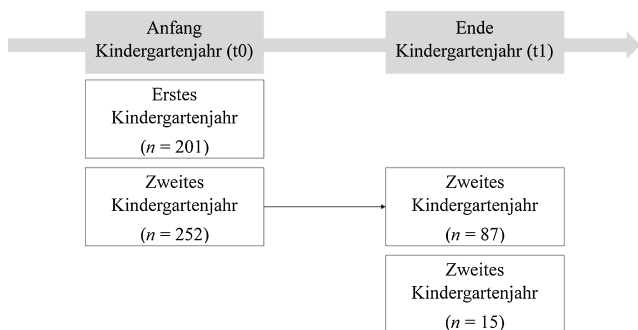


Abbildung 1. Untersuchungsdesign

Instrument: MARKO-D

Das Diagnoseinstrument MARKO-D umfasst 55 Items und baut auf dem im Abschnitt *Entwicklungsmodell arithmetischer Konzepte* beschriebenen Modell von Fritz und Ricken auf (Fritz et al., 2013; Fritz et al., 2018; Ricken et al., 2013). Die Items sind fünf Niveaus zugeteilt. Die Abgrenzung der Niveaus wurde mit dem Raschmodell anhand der Normstichprobe empirisch überprüft. Dies erfolgte unter anderem mit dem sogenannten Item-Separation-Index (Wright & Masters, 1982). Mit diesem wenig bekannten Maß wird geschätzt, welche maximale Anzahl von statistisch trennbaren *Teilverteilungen* (Strata) der Items grundsätzlich möglich ist (Details und Formel siehe Abschnitt *Datenanalyse*). Der Index ergab einen Wert von 3.26, was vier bis fünf unterschiedliche Niveaus nahelegt (Ricken et al., 2013).

Die Normierungsstichprobe von MARKO-D besteht aus 1 095 Kindern (48 % Mädchen) aus Deutschland im Alter zwischen 4;0 und 7;3 Jahren. Für jedes Halbjahr liegen Altersnormwerte vor, wobei die Größe der Normgruppen variiert: bis einschließlich 4;5 Jahre ($n = 84$), 4;6 bis 4;11 Jahre ($n = 179$), 5;0 bis 5;5 Jahre ($n = 305$), 5;6 bis 5;11 Jahre ($n = 329$), 6;0 bis 6;5 Jahre ($n = 172$) sowie 6;6 Jahre und älter ($n = 26$). Der Einzeltest dauert zwischen 20 und 30 Minuten und die Items werden im Rahmen einer Geschichte um zwei Eichhörnchen präsentiert.

Die Reliabilität (WLE-Personenschätzwert) von MARKO-D beträgt .91, die Re-Test-Reliabilität .89 (zweite Messung nach zwei Wochen). Die *unweighted Mean Square* Werte (MNSQ) von allen 55 Items lagen zwischen 0.70 und 1.30. Das heißt, die Antworten zu allen Items wurden durch das Raschmodell ausreichend vorhergesagt.

Stichprobe

In die Stichprobe einbezogen wurden insgesamt 468 Kinder aus 28 Kindergartengruppen der deutschsprachigen Schweiz, 201 Kinder am Anfang des ersten Kindergartenjahres (47 % Mädchen; Alter Monate $M = 57,4$, $SD = 4.8$) und 252 Kinder am Anfang des zweiten Kindergartenjahres (48 % Mädchen; Alter Monate $M = 69.6$, $SD = 4.0$). Somit nahmen 453 Kinder beim ersten Messzeitpunkt (t_0) teil. Beim zweiten Messzeitpunkt (t_1) konnten 87 Kinder im zweiten Kindergartenjahr für eine erneute Teilnahme am Ende des Kindergartenjahres gewonnen werden. Zusätzlich nahmen 15 Kinder zum ersten Mal teil (Neuzugänge). Somit wurden bei t_1 102 Kinder am Ende des zweiten Kindergartenjahres (57 % Mädchen; Alter Monate $M = 77.2$, $SD = 4.6$) aus 13 Kindergartengruppen getestet. Eine Drop-Out-Analyse ergab keine signifikanten Unterschiede in der Testleistung zum Messzeitpunkt t_0 zwischen den Kindern im zweiten Kindergartenjahr, die nur zum Messzeitpunkt t_0 teilnahmen ($n = 164$) und jenen, die zu beiden Messzeitpunkten t_0 und t_1 teilnahmen ($n = 87$). Somit kann eine systematische Verzerrung der Stichprobe durch Ausfälle ausgeschlossen werden.

Datenanalyse

Zur Überprüfung der Eignung des Instruments für die Schweizer Stichprobe wurden die Daten mittels des eindimensionalen Raschmodells, das zur Item-Response-Theorie (IRT) gehört, mit dem R-Package *TAM* analysiert (Robitzsch, Kiefer & Wu, 2018). Im Raschmodell werden Testergebnisse als Indikatoren latenter Dimensionen betrachtet und es wird die Zustimmung zu Items in Abhängigkeit von der Merkmalsausprägung der untersuchten Person modelliert (Rost, 2004; Bortz & Döring, 2006). Personen mit demselben Personen-Fähigkeitsparameter weisen die gleiche Fähigkeit auf, unabhängig von den Items, die sie bearbeitet haben. Umgekehrt werden Items mit derselben Schwierigkeit von Personen mit denselben Fähigkeits-Parametern mit vergleichbarer Lösungswahrscheinlichkeit gelöst (Koller, Alexandrowicz & Hatzinger, 2012). Die Personenfähigkeits-Parameter werden mit dem *Weighted Likelihood Estimate* (WLE) nach Warm (1989) geschätzt. Das heißt, dass die Testleistung der einzelnen

Kinder in Abhängigkeit von der Itemschwierigkeit und messfehlerkorrigiert ermittelt wird. Ein wichtiger Kennwert der Items im Raschmodell ist der *Infit-MNSQ-Wert*, der Aufschluss darüber gibt, wie gut einzelne Items ins Modell passen. Items mit einem Wert von 1 passen perfekt ins Messmodell, Werte über 1 weisen auf einen Mangel an Vorhersagbarkeit der Antworten aus dem Personenschätzwert hin, Werte unter 1 auf eine zu hohe Vorhersagbarkeit. Dies erlaubt den Ausschluss von ungeeigneten Items, die vom Erwartungswert abweichen. Je nach Quelle werden unterschiedliche Toleranzgrenzen der MSNQ-Werte vorgeschlagen. Gemäß E. V. Smith und R. M. Smith (2004) können Werte zwischen 0.70 und 1.30 akzeptiert werden. Diese Grenzwerte wurden – wie auch in MARKO-D – für die vorliegende Studie übernommen.

Zur Überprüfung der Niveaubildung wurde analog zum MARKO-D die Berechnung des *Item-Separation-Index G* vorgenommen (Bond & Fox, 2015; Wright & Masters, 1982). Diese Maßzahl basiert auf der Item Separation Reliabilität der Raschskala und wird mittels folgender Formel berechnet:

$$G = \sqrt{R/(1-R)}. \quad (1)$$

G beziffert das Verhältnis der wahren Standardabweichung der Itemmesswerte relativ zu deren Fehler. Ausgehend davon kann die Anzahl der systematisch trennbaren Schwierigkeitsniveaus H bestimmt werden.

$$H = (4G+1)/3. \quad (2)$$

H spiegelt die Obergrenze der Anzahl Item-Strata wider, deren Zentren statistisch überzufällig auseinanderliegen können. Die so definierte maximale Anzahl unterscheidbarer Niveaus von Itemschwierigkeiten stellt demnach eine Funktion der Reliabilität der Raschskala dar. Liegt die Reliabilität hoch (Testaufgaben liefern für alle Zonen der Skala mit dem Raschmodell konsistente, genaue Information), erlaubt dies eine feinere Unterscheidung bzw. eine größere Anzahl von systematisch reproduzierbaren Schwierigkeitsniveaus als bei einer niedrigen Reliabilität.

In einem weiteren Schritt wurde die Eignung der Altersnormen des MARKO-D für die Schweizer Stichprobe bzw. den Zusammenhang zwischen der Kindergartenerfahrung und den Testergebnissen untersucht. Hierfür erfolgten Gruppenvergleiche in Bezug auf Alter und Kindergartenerfahrung mittels einfaktorieller Varianzanalyse und *t*-Tests.

Zur Überprüfung der zeitbezogenen Messinvarianz wurde zunächst die Veränderung der Leistung der Teilstichprobe von t0 auf t1 mit einem *t*-Test für abhängige Stichproben auf ihre Signifikanz überprüft. Anschließend

folgten die Analysen mittels *Differential-Item-Functioning* (DIF) auf der Ebene der Items mit der Teilstichprobe der 87 Kinder, die zu beiden Messzeitpunkten teilnahmen (Schwab & Helm, 2015). Hierbei wird die Lage der Schwierigkeitsparameter der Items über die Zeit analysiert. Im Idealfall ist die Schwierigkeit aller Aufgaben (in Logits) über die Testzeitpunkte hinweg vollständig parallel. Es wurde ein Vorgehen mit virtuellen Personen gewählt, bei dem die zeitpunktspezifischen Datenmatrizen vertikal an die Datenmatrix des ersten Testzeitpunktes angehängt werden. Das heißt, dass zwei Messungen von derselben Person als Messungen von zwei unterschiedlichen Personen betrachtet werden. So kann für die DIF-Analyse der Testzeitpunkt als Gruppierungsvariable eingesetzt werden (Hartig & Kühnbach, 2006). Das heißt, dass die aufgabenbezogene Differenz zwischen den Messzeitpunkten der Differenz der mittleren Aufgabenschwierigkeit des ganzen Tests entsprechen sollte. Itembezogene Abweichungen der Parameter werden dabei ausgehend vom Ideal einer Nulldifferenz berechnet. Bei vollständiger Parallelität der Aufgabenschwierigkeiten müsste die Parameterdifferenz nach Abzug des insgesamt bestehenden Offsets für jedes Parameterpaar 0 betragen. Zur Interpretation können die Grenzwerte nach Paek und Wilson (2011) herangezogen werden. Differenzen bis maximal 0.426 gelten bei nicht gegebener Signifikanz ($p \geq .05$) als vernachlässigbar. Der Grad an DIF gilt als mittel, wenn $p < .05$ und die Parameterdifferenz kleiner als 0.638 ist. Eine DIF gilt als groß, wenn die Differenz (bei $p < .05$) größer als 0.638 ist.

Ergebnisse

Itemanalysen und Niveauüberprüfung

Für die Itemanalysen mit dem Raschmodell wurden alle 55 Items des MARKO-D-Tests und die Daten der Gesamtstichprobe ($N = 555$) verwendet. In der Gesamtstichprobe werden die beiden Messungen der 87 Kinder, die den Test zwei Mal bearbeitet haben (am Anfang des zweiten Kindergartenjahres und am Ende des zweiten Kindergartenjahres) als separate Fälle behandelt. Da die Testzeitpunkte mindestens ein halbes Jahr auseinanderliegen, können Re-Testeffekte ausgeschlossen werden. Zudem führte eine Analyse, in der nur die Daten einer Messung berücksichtigt wurden (mit $n = 468$) zu nahezu identischen Ergebnissen.

Die Item Separation Reliabilität des Tests für die Schweizer Stichprobe war hoch (.99). Der Infit-MNSQ von 52 Items lag im Toleranzbereich 0.71 bis 1.30 und entsprach somit den von E. V. Smith und R. M. Smith

(2004) festgelegten Werten. Anders als in der deutschen Normierungsstichprobe gab es drei Items, deren Fit außerhalb dieser Grenzen lag (siehe Tabelle 1).

In Tabelle 1 sind die Items nach Schwierigkeit (in Logits) für die Schweizer Stichprobe geordnet. Items, die einen schlechten Item-Fit aufweisen, sind kursiv markiert. Fett markierte Linien bezeichnen die Niveaugrenzen gemäß dem Manual des MARKO-D. Mit einer dickeren Linie markiert sind Stellen in der Abfolge der Items, bei denen in der Schweizer Stichprobe besonders deutliche Schwierigkeitsunterschiede festzustellen sind. Dies betraf zwei Stellen an denen sich eine Differenz von mehr als 0.5 Logits zeigte. Fett markiert sind Items, bei denen die Zuordnung zu den Niveaus von der Stichprobe aus Deutschland abweicht. Bei 49 Items stimmte die Niveauezuteilung der Schweizer Stichprobe mit der Normierungsstichprobe aus Deutschland überein, bei sechs Items ergaben sich Abweichungen. Items 15, 29 und 31 lagen in der Schweizer Stichprobe auf einem höheren Niveau als in der deutschen Stichprobe. Die nicht passenden Zuordnungen betrafen Items von Niveau II bis IV.

Für eine weitere empirische Überprüfung der Niveaus wurde analog zum MARKO-D Test der Item-Separation-Index berechnet, dieser betrug $G = 15.78$. Dieser Indexwert entspricht 21 bis 22 unterschiedlichen Niveaus, also drei bis fünf Mal mehr als bei der deutschen Normstichprobe, die einen Item Separation Index von 3.26 aufwies.

Eignung der MARKO-D Altersnormen und Bedeutung der Kindergartenerfahrung

Um die Passung der Deutschen Normen auf die Daten der Schweizer Stichprobe zu überprüfen, wurden die Kinder in die in MARKO-D festgelegten Altersgruppen eingeteilt. In Tabelle 2 sind die Testleistungen der Schweizer Kinder geordnet nach den Altersgruppen von MARKO-D und der Kindergartenerfahrung dargestellt. Hier wird die Bedeutung der Berücksichtigung des Kindergartenbesuchs ersichtlich. So waren bei den Erhebungen ungefähr 55 % der Kinder im Alter von 5;0 bis 5;5 Jahren im ersten Kindergartenjahr und ungefähr 45 % dieser Altersgruppe bereits im zweiten Kindergartenjahr. Bei den Kindern im Alter von 5;6 bis 5;11 und 6;0 bis 6;5 Jahren bestehen ebenfalls Unterschiede bezüglich ihrer Kindergartenerfahrung, diese sind in allen drei untersuchten Stichprobengruppen (Anfang erstes und zweites Kindergartenjahr, Ende zweites Kindergartenjahr) vertreten. Innerhalb der Altersgruppen sind Unterschiede bzw. ein Zuwachs der durchschnittlichen Rohwertsummen in Abhängigkeit des Kindergartenjahres feststellbar. Bei den Kindern im Alter von 5;0 bis 5;5 Jahren gibt es beispielsweise bezüglich der durchschnittlichen Rohwertsumme einen Unterschied von

8 Punkten zwischen dem Anfang des ersten Kindergartenjahres und dem Anfang des zweiten Kindergartenjahres.

Als nächstes wurde der Zusammenhang zwischen der Förderung im Kindergarten in der Schweiz und den mathematischen Kompetenzen der Kinder analysiert. Hierfür wurden die Daten getrennt für die Kindergartenhalbjahre ausgewertet. Insgesamt lösten Kinder am Anfang des ersten Kindergartenjahres im Durchschnitt 34 % der 55 Items richtig ($SD = .19$; Min = .00; Max = .89), Kinder am Anfang des zweiten Kindergartenjahres rund 56 % ($SD = .19$; Min = .07; Max = 1.00) und Kinder am Ende des zweiten Kindergartenjahres rund 66 % ($SD = .19$; Min = .11; Max = 1.00). Mehr als die Hälfte der Kinder am Anfang des ersten Kindergartenjahres lösten alle einfachen Aufgaben. Am Anfang des zweiten Kindergartenjahres bearbeiteten 50 % der Kinder zudem alle Aufgaben mit einer mittleren Schwierigkeit. Am Ende des zweiten Kindergartenjahres lösten mehr als die Hälfte der Kinder alle einfachen und mittelschweren Aufgaben. Eine Signifikanzüberprüfung mittels einfaktorierter Varianzanalyse mit den WLE-Werten ergab, dass sich die Leistung der Kinder in Abhängigkeit der Kindergartenerfahrung signifikant unterschied, $F(2,55) = 118.03$, $p < .001$, bei einem starken Effekt von $f = .65$ nach Cohen (1992).

Berücksichtigt werden muss jedoch nicht nur der Kindergartenbesuch, sondern auch das Alter, da die Auswertung von MARKO-D nach Altersnormen erfolgt. Mit weiteren Analysen wurden deshalb die Leistungen von Kindern gleichen Alters untersucht, die den Kindergarten unterschiedlich lang besucht hatten. Von den Kindern im Alter von 5;0 und 5;5 Jahren waren $n = 61$ zum Zeitpunkt der Datenerhebung im ersten Kindergartenjahr, 49 Kinder dieser Altersgruppe besuchte den Kindergarten zu diesem Zeitpunkt schon das zweite Jahr. Ein Vergleich der WLE-Werte der beiden Gruppen bot sich deshalb an, um die Frage nach einem allfälligen Zusammenhang mit dem Kindergartenbesuch zu untersuchen. Kinder ohne Kindergartenerfahrung wiesen signifikant niedrigere Leistungen auf ($M = -.97$, $SD = 1.66$, $n = 61$), als die Kinder gleichen Alters am Anfang des zweiten Kindergartenjahres ($M = .33$, $SD = 1.32$, $n = 49$) beziehungsweise mit einem Jahr Kindergartenerfahrung, $t(108) = -4.47$, $p < .001$. Die Effektstärke $r = .39$ entspricht einem mittleren Effekt.

Zeitliche Messinvarianz

Die WLE-Reliabilität der 55 Items für die Teilstichprobe von 87 Kindern war hoch und betrug zum ersten Messzeitpunkt (t_0) am Anfang des zweiten Kindergartenjahres 0.94 und zum zweiten Messzeitpunkt (t_1) am Ende des zweiten Kindergartenjahres 0.91. Bei t_0 erreichten die

Tabelle 1. Itemkennwerte für die Gesamtstichprobe (N = 555) geordnet nach Itemschwierigkeit

Niveau	Nummer	Beschreibung	Logits	SE	Infit	DIF	SE
V	26	Wie viele sind es weniger? (5)	3.68	.18	0.97	-0.16	.16
	46	8 Chips legen, 2 blaue mehr als rote	3.40	.17	1.08	-0.03	.16
	33	Zahl um 3 kleiner als 7	3.35	.17	0.94	0.5**	.15
	22	Wie viele Erdbeeren sind es weniger? (2)	3.29	.17	0.91	0.02	.15
	24	Wie viele Eier sind es mehr? (3)	3.07	.16	0.86	0.18	.15
	34	Zahl 2 kleiner als 5	2.71	.15	0.94	0.55***	.14
	39	Rückwärts zählen ab 6 in Zweierschritten	2.57	.14	0.86	-0.05	.14
	32	Zahl um 2 größer als 3	2.48	.14	0.95	0.12	.14
	35	Zahl um 2 größer als 4	2.39	.14	0.93	0.04	.14
	20	Wie viele Sterne sind es mehr? (3)	2.18	.13	1.00	0.04	.14
	38	Vorwärts zählen ab 5 in Zweierschritten	2.16	.13	0.91	-0.03	.14
IV	49	3 Blumen (Chips) weggenommen ergibt 5, wie viele waren vorher da?	1.63	.12	1.05	0.07	.14
	31	Zahl um 1 kleiner als 9	1.58	.12	0.84	0.39**	.14
	45	6 Chips legen, mehr blau als rot	1.46	.12	0.83	0.26	.14
	44	9 Chips legen, davon 4 blau	1.41	.12	0.97	0.07	.14
	29	Zahl um 1 kleiner als 5	1.35	.12	1.09	0.19	.14
	47	4 Blumen (Chips), wie viele fehlen bis 10?	1.23	.12	1.14	0.08	.13
	43	5 Chips legen, davon 3 rot	1.13	.11	0.88	0.31*	.14
III	30	Zahl um 1 größer als 7	0.91	.11	0.99	0.08	.13
	18	Reihenfolge ergänzen (5)	0.89	.11	1.02	-0.06	.13
	7	Nachfrage zu Item 6: Wie viele sind es? Nachfrage: Und wenn du hier anfängst?	0.61	.11	1.33	-0.31*	.13
	54	Gleich mächtige Menge mit Chips legen (4)	0.58	.11	0.95	0.2	.13
	17	Reihenfolge ergänzen (2)	0.56	.11	1.04	-0.19	.13
	16	Reihenfolge ergänzen (5)	0.50	.11	1.21	-0.26	.13
	12	Im Kopf: 5 Nüsse, 3 gegessen. Wie viele bleiben?	0.48	.11	1.06	0.24	.14
	36	4 Sterne, 3 Sterne verdeckt. Wie viele zusammen?	0.44	.11	1.04	0.04	.13
	14	Im Kopf: 3 Nüsse von mir, 3 vom Bruder. Wie viele?	0.37	.11	0.87	-0.06	.14
	55	Gleich mächtige Menge mit Chips legen (5)	0.33	.11	0.94	0.43**	.14
	13	Aufgabe von Item 12 (5-3) mit Chips legen	0.27	.11	0.89	0.55***	.14
	II	5	Zahl zwischen 5 und 7	0.20	.11	0.86	0.04
10		Im Kopf: 2 Nüsse, noch zwei geschenkt. Wie viele?	0.18	.11	1.05	0.04	.14
37		5 Sterne, 3 Sterne verdeckt. Wie viele zusammen?	0.16	.11	0.90	0.15	.14
11		Aufgabe von Item 10 (2 + 2) mit Chips legen	-0.05	.11	1.03	0.31*	.14
4		Zahl zwischen 2 und 4	-0.08	.11	0.87	-0.24	.14
15		Aufgabe von Item 14 (3 + 3) mit Chips legen	-0.23	.11	0.87	0.24	.14
2		Zahl vor 5	-0.24	.11	0.95	-0.72***	.13
28		6 Hasen und 7 Blumen je in einer Reihe. Wo sind mehr?	-0.48	.11	1.31	0.14	.14
3		Zahl nach 5	-0.87	.11	0.94	-0.33*	.14
48		8 Blumen, wie viele wegnehmen bis 6?	-0.91	.11	1.08	0.17	.14
I	27	6 Zapfen und 5 Eicheln je in einer Reihe. Wo sind weniger?	-1.46	.12	1.40	-0.23	.14
	53	Gleich mächtige Mengen mit Chips legen (5)	-1.85	.13	1.07	0.13	.15
	9	Nachfrage zu Item 8: Wie viele sind es? (6)	-2.01	.13	1.00	-0.13	.16

Tabelle 1. Itemkennwerte für die Gesamtstichprobe ($N = 555$) geordnet nach Itemschwierigkeit (Fortsetzung)

Niveau	Nummer	Beschreibung	Logits	SE	Infit	DIF	SE
	41	6 Chips abzählen	-2.16	.13	0.85	-0.51**	.16
	42	5 Chips abzählen	-2.42	.14	0.81	-0.4*	.17
	8	6 Nüsse zählen	-2.56	.14	0.95	-0.42*	.16
	6	9 im Kreis angeordnete Nüsse zählen	-2.58	.14	1.19	0.03	.16
	40	4 Chips abzählen	-2.92	.16	0.89	-0.31	.16
	52	Gleich mächtige Mengen mit Chips legen (4)	-3.02	.16	0.83	0.56**	.17
	1	Zählen bis 10	-3.34	.17	0.81	-0.32	.18
	50	10 Chips aufteilen: jeder 5	-3.49	.18	0.91	1.81***	.19
	25	4 und 9 Pilze in zwei Reihen. Wo sind weniger?	-3.65	.19	0.95	-0.25	.19
	51	8 Chips aufteilen: jeder 4	-3.72	.19	0.90	2.45***	.19
	21	4 und 6 Erdbeeren in zwei Reihen. Wo sind weniger?	-3.72	.19	1.02	-0.74***	.19
	19	3 und 6 Sterne in zwei Reihen. Wo sind mehr?	-3.83	.20	0.94	-0.74***	.19
	23	4 Nester und 7 Eier je in einer Reihe. Wo sind mehr?	-3.98	.20	1.01	-3.93***	.18

Anmerkungen: a) Items mit leichten Underfits (Wert über 1.3) sind kursiv. b) Items, die nicht im Entwicklungsniveau von MARKO-D liegen sind fett gedruckt. c) Bei einer dickeren Linie zwischen zwei Zeilen ist der Schwierigkeitsunterschied zwischen den zwei Items > 0.5 Logits. d) Bei Items mit einer mittleren bis großen DIF ist das Signifikanzniveau angegeben (* $p < .05$, ** $p < .01$, *** $p < .001$).

Tabelle 2. Beschreibung der Untersuchungsstichprobe

Altersgruppe	n (% Mädchen)	Anfang 1. Jahr		Anfang 2. Jahr		Ende 2. Jahr	
		n	M (SD)	n	M (SD)	n	M (SD)
$\leq 4;5$	50 (56)	50	13.2 (8.4)				
4;6 – 4;11	80 (45)	80	20 (10.9)				
5;0 – 5;5	110 (42.7)	61	20.9 (10.8)	49	29.3 (9.3)		
5;6 – 5;11	132 (48.5)	9	22 (7.6)	122	30.2 (10.6)	1	15
6;0 – 6;5	81 (50.6)	1	19	74	33.2 (10.2)	6	39.5 (15.8)
$\geq 6;6$	15 (33.3)			7	28.4 (12.2)	8	26.8 (10.1)
Insgesamt	468 (47.2)	201	18.6 (10.6)	252	30.9 (10.3)	15	31.1 (14.1)

Anmerkung: In dieser Tabelle sind nur die durchschnittlichen Rohwertsummen der ersten Testdurchführung mit den Kindern enthalten, d. h. die Rohwertsummen der Messwiederholung wurden hier nicht berücksichtigt.

Kinder einen durchschnittlichen Rohwert von $M = 29.63$ Punkten ($SD = 11.12$), was einem durchschnittlichen WLE-Personenfähigkeitsparameter von $M = -.55$ ($SD = 1.70$) entspricht. Bei t1 lag der Mittelwert bei $M = 37.25$ Punkten ($SD = 9.82$) bzw. bei einem WLE-Personenfähigkeitsparameter von $M = .56$ ($SD = 1.52$). Die Leistung der Kinder nahm von t0 auf t1 signifikant zu, $t(86) = -11.83$, $p < .001$, mit einem starken Effekt von $r = .79$.

Die Überprüfung der zeitlichen Messinvarianz mittels DIF-Analysen ergab, dass von den insgesamt 55 Items 17 eine große DIF (Differenz > 0.426 , $p < .05$) zeigten. Zwei weitere Items (7 und 11) wiesen eine mittlere DIF auf (Differenz < 0.638 , $p < .05$; vgl. Tabelle 1). Für die Schweizer Stichprobe scheinen sich also 36 von 55 Items für eine Messwiederholung im zweiten Kindergartenjahr mit ei-

nem Abstand von ungefähr einem halben Jahr zwischen den beiden Erhebungen zu eignen.

Diskussion

Mit MARKO-D liegt ein theoretisch fundiertes und empirisch validiertes Instrument zur Erfassung der numerischen Kompetenzen im Vorschulalter vor, das an einer Stichprobe von Kindern aus Deutschland normiert worden ist. Da es keine vergleichbaren Tests mit Schweizer Normen gibt, wurde erstens untersucht, ob das Instrument in einer Schweizer Stichprobe dieselbe Schwierigkeit aufweist wie in der Normierungsstichprobe und ob

die Einteilung in Entwicklungsniveaus übernommen werden kann. Zweitens wurde analysiert, welchen Einfluss die Kindergartenerfahrung auf die mathematische Entwicklung hat und ob die Altersnormen zielführend sind. Drittens wurde überprüft, ob sich der Test für Messwiederholungen über ein Kindergartenjahr hinweg und sich damit für die Erfassung von Lernfortschritten im letzten Kindergartenjahr eignet.

Die Ergebnisse zeigen insgesamt, dass der Test auch in der Schweizer Stichprobe reliabel misst und die Items mit wenigen Ausnahmen den Anforderungen genügen. Insgesamt wiesen die Items in der Schweizer Stichprobe eine ähnliche Schwierigkeit auf wie für die deutsche Stichprobe. Items zum Vergleich und zur Einschätzung von Mengen waren für die Kinder der beiden Stichproben am einfachsten. Für beide Stichproben waren die Items am schwierigsten, bei denen es um das Verständnis der Relation zwischen Mengen und Zahlen geht. Dies entspricht dem Entwicklungsmodell von Fritz et al. (2018). Am Ende des zweiten Kindergartenjahres scheint der Test jedoch zu einfach zu sein, da mehr als die Hälfte der Kinder alle einfachen und mittelschweren Aufgaben richtig lösten.

Allerdings ergeben sich Fragen zur Niveaubildung und damit verbunden zur niveaubasierten Interpretation der Testergebnisse. Die Überprüfung der Anzahl der Niveaus mit dem Item Separation Index ergab in der Schweizer Stichprobe bis zu fünf Mal mehr Niveaus als bei der deutschen Stichprobe. Der Index zeigt jedoch nur an, wie viele Niveaus es aufgrund der gegebenen Reliabilität geben könnte. Vor diesem Hintergrund stellt sich die Frage nach der Gültigkeit eines solchen Wertes für die empirische Bestätigung von theoretisch begründeten Entwicklungsniveaus. Werden die hierarchisch geordneten Items betrachtet, dann zeichnen sich in den Schweizer Daten eher drei Bereiche ab (siehe Tabelle 1). Der erste Bereich (einfachste Items) stimmt mit dem Niveau I (Zählzahl) von MARKO-D überein, der dritte Bereich entspricht in hohem Maß dem Niveau V (Relationalität). Der mittlere Bereich enthält Items aus den Niveaus II bis IV von MARKO-D. Eine Überlappung von Niveau II und III hat sich auch in der deutschen Stichprobe gezeigt (Ricken et al., 2013). Das heißt, dass sich für die Schweizer Stichprobe empirisch drei Bereiche ergeben, die sich auch theoretisch sinnvoll begründen lassen. Es muss aber auch berücksichtigt werden, dass Items aus demselben Subtest, jedoch mit unterschiedlichen Zahlen, zum Teil auf unterschiedlichen Niveaus liegen. Das Item 29 *Welche Zahl ist um 1 kleiner als 5* liegt gemäß dem Manual auf Niveau IV (Relationalität), während das Item 34 *Welche Zahl ist um 2 kleiner als 5* auf Niveau V (Enthalten sein und Inklusion) liegt. Diese Schwierigkeitsunterschiede sind eher auf die Wahl der Zahlen (Zahl um 1 bzw. um 2 größer) zurückzuführen als auf ein unterschiedliches konzeptuelles Ver-

ständnis. Zusammenfassend muss die Frage gestellt werden, ob angesichts dieser Ergebnisse eine Auswertung nach fünf Niveaus sinnvoll ist (Bruckner-Feld, Kastner-Koller & Deimann, 2015).

In Bezug auf die Eignung von Altersnormen konnte aufgezeigt werden, dass nicht nur das Alter, sondern auch der Kindergartenbesuch eine Rolle bei der Entwicklung mathematischer Kompetenzen zu spielen scheint. Gleichaltrige Schweizer Kinder im Alter von fünf Jahren zeigten unterschiedliche Testleistungen, abhängig davon, ob sie den Kindergarten schon besucht hatten oder nicht. Dieses Ergebnis stimmt mit einer Studie überein, die darauf hinweist, dass der mathematische Leistungszuwachs von Kindergartenkindern in der Schweiz unter anderem durch dessen starke Bildungsorientierung bzw. die damit verbundenen Fördermaßnahmen bestimmt sein könnte (Kuratli Geeler, 2019). Das heißt, dass für die Schweiz eine Auswertung nach Kindergartenhalbjahr – unter Berücksichtigung des Alters – gemacht bzw. entsprechende Normentabellen erstellt werden müssten, wie z.B. im Test TEDI-MATH (Kaufmann et al., 2009) oder im MBK 0 (Krajewski, 2018). Auf einer übergeordneten Ebene bedeutet das Ergebnis, dass schulsystembedingte Unterschiede (z.B. Konzept des Kindergartens, Einschulungsalter, Selektionszeitpunkte) bei der Normierung von Tests berücksichtigt werden müssen (Moser Opitz, Stöckli, Grob, Reusser & Nührenbörger, 2020).

Eine weitere Zielsetzung der Studie bestand darin, die Eignung des Tests für längsschnittliche Messungen zu analysieren. Die Überprüfung der zeitbezogenen Messinvarianz ergab, dass sich ungefähr ein Drittel der Items weniger für eine Messwiederholung im zweiten Kindergartenjahr eigneten. Im hier vorliegenden Kontext der vorschulischen mathematischen Kompetenzen ist zu berücksichtigen, dass bei allgemein schneller Entwicklung die Wahrscheinlichkeit größer ist, dass sich bestimmte Teilkompetenzen rascher entwickeln als andere und somit auch die Wahrscheinlichkeit für Messinvarianz steigt. Zudem muss einschränkend erwähnt werden, dass die Analyse zur Messinvarianz nur mit einer kleinen Teilstichprobe von Kindern im zweiten Kindergartenjahr durchgeführt wurde, und dass das vergleichsweise kleine Sample eher zu größeren, nicht generalisierbaren Ausschlägen von Itemparametern führen kann. Allerdings waren vor allem sehr einfache Items vom Niveau I, die auch in der deutschen Normstichprobe von vielen Kindern richtig gelöst wurden, von DIF betroffen. Das führt zur Vermutung, dass sich Messvarianz für diese Items auch in einer größeren Stichprobe zeigen könnte und weist darauf hin, dass der Test für Kinder im zweiten Kindergartenjahr eher zu viele einfache Items enthält. Das zeigt einmal mehr die Herausforderung, numerische Kompetenzen in dieser Altersstufe längsschnittlich mit demselben Instrument zu messen.

Mit dem MARKO-D steht im deutschsprachigen Raum ein valides Instrument zur Erhebung von Mathematik- und Rechenkonzepten im Vorschulalter zur Verfügung, das für Schweizer Kindergartenkinder geeignet ist. Am Ende des zweiten Kindergartenjahres scheint er jedoch zu einfach zu sein. Des Weiteren muss die ausschließliche Verwendung von Altersnormen hinterfragt werden, weil der Einfluss der Förderung im Kindergarten nicht berücksichtigt wird. Die Auswertung und die Normierung sollte demnach nach Alter und nach Kindergartenhalbjahr erfolgen. Aufgrund der hohen Anzahl von Items mit großer DIF und Items aus Niveau I stellt sich die Frage, ob der Test für längsschnittliche Messungen eingesetzt werden kann. Eine Auswahl an Items – eventuell erweitert mit zusätzlichen und schwierigeren Items – könnte dafür geeignet sein.

Literatur

- Anders, Y. (2012). *Modelle professioneller Kompetenzen für frühpädagogische Fachkräfte. Aktueller Stand und ihr Bezug zur Professionalisierung. Expertise zum Gutachten „Professionalisierung in der Frühpädagogik“*. München: Knoblingdesign
- Aunio, P., Heiskari, P., Van Luit, J. E. & Vuorio, J.-M. (2015). The development of early numeracy skills in kindergarten in low-, average- and high-performance groups. *Journal of Early Childhood Research*, 13(1), 3–16. <https://doi.org/10.1177/1476718X14538722>
- Aunio, P. & Räsänen, P. (2016). Core numerical skills for learning mathematics in children aged five to eight years – a working model for educators. *European Early Childhood Education Research Journal*, 24, 684–704. <https://doi.org/10.1080/1350293X.2014.996424>
- Bond, T. & Fox, C. M. (2015). *Applying the Rasch Model. Fundamental measurement in the human sciences*. New York, NY: Routledge.
- Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer.
- Bruckner-Feld, J., Kastner-Koller, U. & Deimann, P. (2015). Testbesprechung: MARKO-D. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 47, 219–223. <https://doi.org/10.1026/0049-8637/a000131>
- Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, 1 (3), 98–101. <https://doi.org/10.1111/1467-8721.ep10768783>
- Deutscheschweizer Erziehungsdirektorenkonferenz (D-EDK). (2014). *Lehrplan 21*. Verfügbar unter: <https://www.lehrplan21.ch>
- Fritz, A., Ehlert, A. & Leutner, D. (2018). Arithmetische Konzepte aus entwicklungspsychologischer Sicht. *Journal für Mathematik-Didaktik*, 39, 7–41. <https://doi.org/10.1007/s13138-018-0131-6>
- Fritz, A., Ehlert, A. & Balzer, L. (2013). Development of mathematical concepts as the basis for an elaborated mathematical understanding. *South African Journal of Childhood Education*, 3(1), 38–67.
- Gallit, F., Wyschkon, A., Poltz, N., Moraske, S., Kucian, K., Aster, M. von et al. (2018). Henne oder Ei: Reziprozität mathematischer Vorläufer und Vorhersage des Rechnens. *Lernen und Lernstörungen*, 7(2), 81–92. <https://doi.org/10.1024/2235-0977/a000205>
- Gasteiger, H., Brunner, E. & Chen, C. S. (2020). Basic conditions of early mathematics education – a comparison between Germany, Taiwan and Switzerland. *International Journal of Science and Mathematics Education*. <https://doi.org/10.1007/s10763-019-10044-x>
- Göbel, S. M., Watson, S. E., Lervåg, A. & Hulme, C. (2014). Children's arithmetic development: It is number knowledge, not the approximate number sense that counts. *Psychological Science*, 25, 789–798. <https://doi.org/10.1177/0956797613516471>
- Hartig, J. & Kühnbach, O. (2006). Schätzung von Veränderung mit „plausible values“ in mehrdimensionalen Rasch-Modellen. In A. Ittel & H. Merckens (Hrsg.), *Veränderungsmessung und Längsschnittstudien in der Erziehungswissenschaft* (S. 27–44). Wiesbaden: Verlag für Sozialwissenschaften.
- Hauser, B., Vogt, F., Stebler, R. & Rechsteiner, R. (2014). Förderung früher mathematischer Kompetenzen: Spielintegriert oder trainingsbasiert. *Frühe Bildung*, 3, 139–145. <https://doi.org/10.1026/2191-9186/a000144>
- Jörns, C., Schuchardt, K., Grube, D. & Mähler, C. (2014). Spielorientierte Förderung numerischer Kompetenzen im Vorschulalter und deren Eignung zur Prävention von Rechenschwierigkeiten. *Empirische Sonderpädagogik*, 6, 243–259.
- Kaufmann, L., Nuerk, H.-Ch., Graf, M., Krinzinger, H., Delazer, M. & Willmes, K. (2009). *TEDI-MATH: Test zur Erfassung numerisch-rechnerischer Fertigkeiten vom Kindergarten bis zur 3. Klasse*. Bern: Huber.
- Koller, I., Alexandrowicz, R. & Hatzinger, R. (2012). *Das Rasch-Modell in der Praxis: Eine Einführung mit eRm*. Wien: facultas wuv.
- Krajewski, K. & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction*, 19, 513–526. <https://doi.org/10.1016/j.learninstruc.2008.10.002>
- Krajewski, K. (2018). *MBK 0: Test mathematischer Basiskompetenzen im Kindergartenalter*. Göttingen: Hogrefe.
- Kuratli Geeler, S. (2019). *Mathematische Kompetenzen von Kindergartenkindern: Überprüfung eines Testinstrumentes und Analyse von Unterschieden in der numerischen Leistungsentwicklung*. Universität Zürich: Philosophische Fakultät. <https://doi.org/10.5167/uzh-171088>
- Langhorst, P., Hildenbrand, C., Ehlert, A., Ricken, G. & Fritz, A. (2013). Mathematische Bildung im Kindergarten – Evaluation des Förderprogramms „Mina und der Maulwurf“ und die Betrachtung von Fortbildungsvarianten. In M. Hasselhorn, A. Heinze & W. Schneider (Hrsg.), *Diagnostik mathematischer Kompetenzen* (Jahrbuch der pädagogisch-psychologischen Diagnostik, Reihe Tests und Trends, Bd. 11, S. 113–134). Göttingen: Hogrefe.
- Lyons, I. M., Bugden, S., Zheng, S., De Jesus, S. & Ansari, D. (2018). Symbolic number skills predict growth in nonsymbolic number skills in kindergarteners. *Developmental Psychology*, 54, 440–457. <https://doi.org/10.1037/dev0000445>
- Missall, K. N., Mercer, S. H., Martinez, R. S. & Casebeer, D. (2012). Concurrent and longitudinal patterns and trends in performance on early numeracy curriculum-based measures in kindergarten through third grade. *Assessment for Effective Intervention*, 37(2), 95–106. <https://doi.org/10.1177/1534508411430322>
- Moser Opitz, E., Stöckli, M., Grob, U., Reusser, L. & Nührenbörger, M. (2020). *BASIS-MATH-G2+: Gruppentest zur Basisdiagnostik Mathematik für das vierte Quartal der 2. Klasse und das erste Quartal der 3. Klasse*. Bern: Hogrefe.
- Paek, I. & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood


- estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, 71, 1023–1046. <https://doi.org/10.1177/0013164411400734>
- Putnick, D. L. & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>
- Ricken, G., Fritz, A. & Balzer, L. (2013). *MARKO-D: Mathematik- und Rechenkonzepte im Vorschulalter – Diagnose*. Göttingen: Hogrefe.
- Robitzsch, A., Kiefer, T. & Wu, M. (2018). *TAM: Test analysis modules*. Retrieved from <https://cran.r-project.org/package=TAM>
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Hans Huber.
- Schneider, W., Küspert, P. & Krajewski, K. (2016). *Die Entwicklung mathematischer Kompetenzen*. Paderborn: Ferdinand Schöningh.
- Schuchardt, K., Piekny, J., Grube, D. & Mähler, C. (2014). Einfluss kognitiver Merkmale und häuslicher Umgebung auf die Entwicklung numerischer Kompetenzen im Vorschulalter. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 46(1), 24–34. <https://doi.org/10.1026/0049-8637/a000099>
- Schwab, S. & Helm, C. (2015). Überprüfung von Messinvarianz mittels CFA und DIF-Analysen. *Empirische Sonderpädagogik*, 7, 175–193.
- Smith, E. V. & Smith, R. M. (2004). *Introduction to Rasch measurement: Theory, models and applications*. Maple Grove, MN: JAM Press.
- Stock, P., Desoete, A. & Roeyers, H. (2007). Early markers for arithmetic difficulties. *Educational & Child Psychology*, 24(2), 28–39.
- Vogt, F., Hauser, B., Stebler, R., Rechsteiner, K. & Urech, C. (2018). Learning through play – pedagogy and learning outcomes in early childhood mathematics. *European Early Childhood Education Research Journal*, 26, 589–603. <https://doi.org/10.1080/1350293X.2018.1487160>
- Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450. <https://doi.org/10.1007/BF02294627>
- Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analyses*. Chicago, IL: MESA Press.

Historie

Onlineveröffentlichung: 05.11.2020

ORCID

Dr. Ariana Garrote

 <https://orcid.org/0000-0001-7432-7872>

Dr. Ariana Garrote

Pädagogischen Hochschule
Fachhochschule Nordwestschweiz
Obere Sternengasse 7
4502 Solothurn
Schweiz
ariana.garrote@fhnw.ch