VII Lesen Kinder, die nicht in der Kita waren, am Ende der Grundschule schlechter?

Isa Steinmann, Laura R. Zieger, Nina Hogrebe & Rolf Strietholt

Während frühkindliche Bildung und Betreuung mit hohen Erwartungen an Lerneffekte verknüpft sind, ist die diesbezügliche Forschungslage bislang nicht eindeutig. Das Ziel der vorliegenden Studie ist es, den Effekt des Nichtbesuchens von Kitas auf die Leseleistung in der vierten Jahrgangsstufe zu untersuchen: Wie hoch ist die durchschnittliche Leseleistung solcher Kinder, die keine Kita besucht haben, im Vergleich dazu, wenn sie solche Institutionen besucht hätten? Propensity Score Matching Methoden werden auf PIRLS Daten angewendet, um ein randomisiertes Experiment zu imitieren und den Effekt für die Messzeitpunkte 2001, 2006 und 2011 zu schätzen. Die Befunde zeigen, dass Kinder, die keine Kita besucht haben, aus bildungsbenachteiligten Familien stammen. Gleichzeitig ist ihre durchschnittliche Leseleistung in der vierten Klasse nicht statistisch signifikant geringer, als die von Kindern mit vergleichbarem Hintergrund, die mehr als zwei Jahre eine Kita besucht haben. Die Ergebnisse der Studie werden unter Berücksichtigung methodischer Limitationen im Hinblick auf ihre politische Bedeutung diskutiert.

Schlüsselwörter: Kita, bildungsbenachteiligte Kinder, PIRLS, Propensity Score Matching, Leseleistung

1 Einleitung und theoretischer Hintergrund

In Deutschland haben Kinder ab dem vollendeten dritten Lebensjahr seit 1996 das Recht, eine Kindertageseinrichtung (Kita) zu besuchen, also eine institutionalisierte Form der Kindertagesbetreuung wahrzunehmen. Mittlerweile nehmen 94.9% aller drei- bis fünfjährigen Kinder in Deutschland an solchen Angeboten teil (DESTATIS, 2015). In wissenschaftlichen, wirtschaftlichen und politischen Debatten wird diese Form der frühkindlichen Bildung und Betreuung zunehmend als bedeutsamer politischer Aufgabenbereich wahrgenommen, was sich auch in dem Ausbau der Kindertagesbetreuung für Kinder unter drei Jahren, der Implementierung der Bil-

dungspläne sowie diskutierten Professionalisierungszielen widerspiegelt. Mit dem Ausbau von Kitas wird nicht nur das Ziel einer besseren Vereinbarkeit von Familie und Beruf verfolgt, sondern auch die Förderung von Vorläuferfähigkeiten schulischer Kompetenzen beispielsweise im Bereich des Spracherwerbs, insbesondere auch bei Kindern mit bildungsbenachteiligten Familienhintergründen (vgl. Hogrebe, 2014; Anders, 2013).

Frühe Kindertagesbetreuung wird grundsätzlich als gute Fördergrundlage für spätere schulische Leistungsentwicklungen angenommen (z. B. Knudsen, Heckman, Cameron & Shonkoff, 2006) und bietet den ersten Eingriff- oder Steuerpunkt für institutionalisierte Bildung. Kitas stellen neben den Familien eine zentrale Lernumgebung für Kinder dar, was innerhalb der ökologischen Theorien der menschlichen Entwicklung auch im Zusammenspiel mit weiteren Faktoren modelliert wird. Nach Bronfenbrenner (1990) ist menschliche Entwicklung das Resultat eines Zusammenspiels von individuellen Voraussetzungen und dem Umfeld, in dem man aufwächst. Das Umfeld besteht hierbei aus voneinander unabhängigen dynamischen Strukturen, die ineinander eingebettet sind und von der distalen Makro- bis zur proximalen Mikroebene reichen. Strukturen auf einer eher distalen Ebene – wie kulturelle Werte und gesellschaftliche Strukturen, aber auch das Arbeitsleben der Eltern – haben keinen direkten Einfluss auf die Entwicklung des Kindes. Da diese jedoch mit dem proximalen Umfeld zusammenhängen, üben sie ihren Einfluss indirekt aus. Je proximaler Umfeldfaktoren sind, desto unmittelbarer wirken sie sich auf die Entwicklung des Kindes aus. Für Bemühungen, sprachliche Fähigkeiten zu fördern, bilden Familien und Kitas demnach die zwei bedeutsamsten Lernumgebungen für junge Kinder.

Gleichzeitig lässt sich beobachten, dass die Kitateilnahme mit Hintergrundmerkmalen der Familien zusammenhängt, wobei Kinder aus benachteiligten Familien seltener Kitas besuchen. Durch eine Erweiterung der ökologischen Theorie der menschlichen Entwicklung können diese Selektionsmechanismen in Begründungszusammenhänge gestellt werden. Gemäß des Personen-Prozess-Kontext-Zeit-Modells werden diese Mechanismen als abhängig von persönlichen Voraussetzungen und Einflüssen aus Umwelt und Zeit definiert (Bronfenbrenner & Morris, 2006). Early und Burchinal (2001) wenden dies auf die Kitateilnahme an und zeigen, dass die Präferenz der Eltern für bestimmte Betreuungseigenschaften (Personen) und ihre daraus folgende Einstellung für oder gegen institutionalisierte frühkindliche Bildung (Mechanismus/Prozess) durch ihre Lebensumstände beeinflusst werden (Kontext). Sie fanden einen Zusammenhang zwischen hohem Einkommen und der Entscheidung für externe Betreuungsformen z. B. durch

Verwandte oder Betreuungsinstitutionen. Gleichzeitig entscheiden sich Eltern mit Migrationshintergrund in dieser Hinsicht eher für die Betreuung durch Verwandte. Es existieren weitere empirische Befunde, nach denen Eltern mit niedrigerem Bildungs- und Beschäftigungsstatus ebenfalls die Betreuung durch Verwandte dem Kitabesuch vorziehen. Häufige Gründe hierfür sind Kosten und Komfort. Insgesamt ist die Wahrscheinlichkeit bei höher gebildeten Eltern größer als bei anderen, dass sie ihr Kind in die Kita schicken. Diese Tendenz resultiert aus lernfokussierten kulturellen Normen, Wissen über Kindertagesbetreuungsangebote und den nötigen Ressourcen wie dem Einkommen und flexiblen Arbeitszeiten, welche durch ein besseres Beschäftigungsverhältnis möglich sind (z. B. Grogan, 2012; Kim & Fram, 2009).

1.1 Empirische Befunde zu den Effekten der Kitateilnahme auf spätere Kompetenzen

Während institutionelle frühkindliche Bildung weltweit expandiert, bestehen hierbei deutliche Unterschiede zwischen Ländern (z. B. UNESCO, 2006). In internationalen Vergleichsstudien wie dem *Programme for International Student Assessment* (PISA) oder der *Progress in International Reading Literacy Study* (PIRLS) werden wiederkehrend (unterschiedlich stark ausgeprägte) Leistungsunterschiede zwischen Kindern mit und ohne vorschulischem Kitabesuch beobachtet. Solche Schüler/-innen, die in den Leistungstests niedrige Punktwerte erzielen, haben überproportional häufig keine frühkindlichen Bildungsinstitutionen besucht.

Beispielsweise zeigt sich in PIRLS 2011 ein positiver Zusammenhang zwischen dem Kitabesuch und der Leseleistung in der vierten Jahrgangsstufe. Darauf basierend schlussfolgern Mullis, Martin, Foy und Drucker (2012), dass die Kindertagesbetreuung entscheidend für den späteren Lesekompetenzerwerb und demnach auch den Schulerfolg sei. Ein aktueller Report basierend auf den Daten von PISA kommt zu ähnlichen Schlüssen (Schleicher, 2014). Vor diesem Hintergrund wird die Kindertagesbetreuung auch als politischer Ansatzpunkt zur zukünftigen Verbesserung von schulischen Kompetenzen angesehen.

Gleichzeitig können korrelative Befunde aus querschnittlich angelegten Studien nur vorsichtig interpretiert werden, da Verzerrungen aufgrund konfundierender Variablen denkbar sind. Wie oben dargelegt, hängt die Kitateilnahme auch von elterlichen Präferenzen und Ressourcen ab. Entsprechende Forschungsbefunde zeigen, dass Kinder aus bildungsbenachteiligten Elternhäusern oftmals nicht oder seltener, beziehungsweise eher an solchen Angeboten mit einer niedrigeren Prozessqualität teilnehmen, als

solche mit einem privilegierten Hintergrund (z. B. Hynes & Habasevich-Brooks, 2008; Pianta, Barnett, Burchinal & Thornburg, 2009). Solche Selektionsmechanismen wirken sich deutlich auf korrelative Befunde aus: Zum Beispiel reduziert sich der starke Zusammenhang zwischen dem Kitabesuch und der späteren Leistung in Mathematik in PISA 2012 in fast allen Ländern erheblich, sobald der sozioökonomische Hintergrund mitberücksichtigt wird (OECD, 2013).

Betrachtet man ausschließlich solche Studien, die einen kausalen Schluss von der Kitateilnahme auf spätere Kompetenzentwicklungen zulassen, sind die Forschungsbefunde nicht so eindeutig, wie manchmal diskutiert (vgl. Duncan & Magnuson, 2013). Ein Grund dafür sind die international unterschiedlichen Arten und Ausrichtungen von institutioneller frühkindlicher Bildung, welche dementsprechend auch zu unterschiedlichen Zusammenhangsausprägungen von Teilnahme und familiären Hintergrundmerkmalen führen. Wenn sich vorschulische Programme beispielsweise spezifisch an Kinder aus bildungsbenachteiligten Familien richten, dann liegen keine zufälligen Stichproben vor. Dies schränkt die Generalisierbarkeit solcher Befunde insbesondere auch in Bezug auf Systeme ein, die universell (d. h. auf alle Kinder einer Altersgruppe) ausgerichtet sind und auf einer freiwilligen Teilnahme basieren, wie beispielsweise in Deutschland.

Demnach verwundert es nicht, dass internationale Literatur- und Meta-analysen zu den Effekten von frühkindlichen Bildungsprogrammen auf spätere Schulleistungen substantielle Effektgrößenunterschiede zwischen verschiedenen Studien aufzeigen. Diese Unterschiede stehen primär im Zusammenhang mit den untersuchten Programmarten und deren Zielgruppen (Barnett, 2011; Burger, 2010; Camilli, Vargas, Ryan & Barnett, 2010; Chambers, Cheung, Slavin, Smith & Laurenzano, 2010; Duncan & Magnuson, 2013; Pianta et al., 2009). Grundsätzlich werden Langzeiteffekte von Kitaprogrammen bereits seit 50 Jahren untersucht. Charakteristisch für solche Programme wie dem *High/Scope Perry Preschool* Projekt (Schweinhart, Montie, Xiang, Barnett, Belfield & Nores, 2005) oder dem *Abecedarian* Programm (Campbell & Ramey, 1995) sind die kleinen Stichprobengrößen von ungefähr 100 bildungsbenachteiligten Kindern und die hohe Intensität und Qualität der Programme, welche sich auch durch elterliche Unterstützungselemente auszeichnen. Experimentalstudien konnten langanhaltende förderliche Effekte dieser Programme auf zahlreiche Outcomes im Schulalter, wie auch die Lesefähigkeit, belegen. Durch die hohe Prozessqualität, die lokale Ausrichtung, die spezifischen Zielgruppen sowie den historischen und politischen Kontext der Studien sind deren Befunde jedoch nur be-

grenzt valide übertragbar, wodurch Schlussfolgerungen für Länder oder Staaten nur schwer zu ziehen sind (vgl. Duncan & Magnuson, 2013).

Aktuellen internationalen Studien zur Effektivität frühkindlicher Förderung von benachteiligten Kindern ist es nicht möglich, die vielversprechenden Resultate dieser kleinformatigen Studien zu replizieren (vgl. Barnett, 2011). Beispielsweise zeigt die *Head Start* Studie keine signifikanten Effekte auf spätere Leistungsmaße (Puma, Bell, Cook, Heid, Broene, Jenkins, Mashburn & Downer, 2012). Ähnlich verhält es sich bei der *Early Head Start* Studie (EHS), bei der zumindest keine allgemeinen förderlichen Einflüsse gefunden wurden. Bezogen auf die Lesekompetenzen konnten jedoch Effekte für solche Kinder gezeigt werden, die weniger ausgeprägte Risikoprofile aufwiesen (Vogel, Xue, Moiduddin, Carlson & Kisker, 2010). Basierend auf den Resultaten der EHS Studie könnte angenommen werden, dass universelle Programme, die sich nicht speziell auf eine Zielgruppe sondern alle Kinder eines Alters fokussieren, eher uneffektiv für bildungsbenachteiligte Kinder sind. Das *Effective Provision of Preschool Education* Projekt (EPPE; 1997–2013), eine einflussreiche europäische Studie zu Systemen mit universeller und freiwilliger Kitateilnahme in England, hat jedoch gezeigt, dass die Kitateilnahme insbesondere auf die am stärksten benachteiligten Kinder einen signifikanten förderlichen Langzeiteinfluss hat. Gleichzeitig unterstreichen die Befunde die Bedeutsamkeit der Qualität der frühkindlichen Lernumgebungen in Systemen mit verpflichtender Kitateilnahme (Sylva, Melhuish, Sammons, Siraj-Blatchford & Taggert, 2008). Die international vergleichende Studie von Hogrebe und Strietholt (2016) macht deutlich, dass es länderspezifische Unterschiede in der Effektivität der frühkindlichen Bildungssysteme zu geben scheint. Nur in zwei von neun in der Untersuchung berücksichtigten Ländern hätten bildungsbenachteiligte Kinder von einem Kitabesuch profitiert. Für Deutschland liegen bisher kaum systematische und methodisch robuste Studien zu den Effekten des Kitabesuchs vor (Anders & Roßbach, 2014).

1.2 Forschungsfrage

Der beschriebene Forschungsstand zeigt uneinheitliche Befunde in Bezug auf die Effektivität von Kindertagesbetreuung auf die Leseleistung am Ende der Primarstufe. Internationale Vergleichsstudien stellen aufgrund ihrer Repräsentativität zwar eine wertvolle Datenquelle dar, gleichzeitig müssen in querschnittlichen Studien Selektionsmechanismen berücksichtigt werden, wenn Effekte der Kitateilnahme auf spätere Leistungsmaße geschätzt werden sollen. In Anlehnung an das Vorgehen von Hogrebe und Strietholt (2016) macht sich die vorliegende Studie Propensity Score Methoden zu-

nutze, um Verzerrungen durch Selektionsmechanismen zu umgehen. Die Studie untersucht, ob Kinder, die keine Kita besucht haben, bessere Lesekompetenzen in der vierten Jahrgangsstufe aufweisen würden, wenn sie Kitaangebote wahrgenommen hätten. Hierzu werden die Daten aus PIRLS 2001, 2006 und 2011 in Deutschland herangezogen, sodass der interessierende Effekt einer fehlenden Kitateilnahme für verschiedene Zeitpunkte untersucht werden kann.

Bevor die inhaltlichen Analysen näher beschrieben werden, folgt eine kurze Einführung in die Anwendung von Propensity Score Methoden zur Schätzung kausaler Effekte.

2 Analysegrundlagen: Propensity Score Methoden und kausale Inferenz

2.1 Kausalität und Selektionseffekte

In vielen Ländern ist der Kitabesuch zwar nicht verpflichtend, aber die meisten Kinder nehmen teil. In der vorliegenden Studie werden die Konsequenzen des fehlenden Besuchs einer Kita für spätere Leseleistungen in Deutschland untersucht. Die Kitateilnahme kann demnach als binäre Variable $A_i = \{0,1\}$ verstanden werden und die Treatmentgruppe besteht aus Kindern, die keine Kita besucht haben. Das Outcome der Leseleistung in der vierten Jahrgangsstufe kann als stetige Variable Y_i dargestellt werden. Unabhängig von der tatsächlichen Gruppenzugehörigkeit bezeichnet Y_{0i} die Leseleistung, wenn das Kind eine Kita besucht hätte ($A_i = 0$), und Y_{1i} die Leseleistung, wenn dieses Kind keine Kita besucht hätte ($A_i = 1$). Die Differenz $Y_{1i} - Y_{0i}$ repräsentiert den kausalen Effekt der Kitateilnahme für das Individuum *i*. Diese Notation wird als Potential Outcome Framework oder als Rubins kausales Modell bezeichnet (Rubin, 1974; Imbens & Rubin, 2015).

Obwohl in der Theorie ein Kind ausreicht, um kausale Effekte zu definieren, sind in der Praxis mindestens zwei Kinder nötig, da es unmöglich ist, beide Leistungswerte an einem Individuum zu beobachten. Deshalb wird das Stichprobenmittel der Leseleistung der Kontrollgruppe $E[Y_{0i} \mid A_i = 0]$ mit dem Stichprobenmittel der Treatmentgruppe $E[Y_{1i} \mid A_i = 1]$ verglichen. Dieser Vergleich lässt jedoch nur dann valide Inferenzen zu, wenn sich andere Prädiktoren der Leseleistung nicht zwischen den Gruppen unterscheiden. Bei der Art der zur Verfügung stehenden Daten ist diese Voraussetzung jedoch im Normalfall nicht gegeben; so bestehen in dem vorliegenden Zusammenhang starke Evidenzen, dass sich teilnehmende und

nicht teilnehmende Kinder in Bezug auf verschiedene Hintergrundmerkmale wie den sozio-ökonomischen Hintergrund unterscheiden (s. o.). Da die spätere Leseleistung von solchen Kovariaten der Kitateilnahme mit beeinflusst wird, schätzt der Gruppenvergleich der Mittelwerte nicht den kausalen Effekt, sondern eine verzerrte Version.

Um hinlänglich beschreiben zu können, wie Propensity Score Methoden mit solchen Selektionsmechanismen umgehen, müssen vorweg zwei Konzepte eingeführt werden: der *Average Treatment Effect* (ATE) und der *Average Treatment Effect on the Treated* (ATT; Imbens, 2004). Der ATE beschreibt den durchschnittlichen Effekt des Treatments für beide Gruppen: die Treatmentgruppe und die Kontrollgruppe. ATT hingegen bezeichnet den durchschnittlichen Effekt des Treatments auf die Treatmentgruppe, also in dieser Studie die Kinder, die keine Kita besucht haben. Diese Unterscheidung ist bei Beobachtungsstudien bedeutsam, da sich die Kontroll- und Treatmentgruppen unterscheiden können. Heckmann und Robb (1986) stellen heraus, dass der ATT besonders für Forschungsfragen relevant ist, die sozialpolitische Entscheidungen betreffen. Im vorliegenden Fall könnte die Politik erwägen, alle Kinder zwischen drei Jahren und der Einschulung zur Kitateilnahme zu verpflichten oder die Steigerung der Teilnahmeraten anderweitig weiter zu forcieren. Dies hätte per se nur Konsequenzen für solche Kinder, die ansonsten *nicht* teilnehmen würden. Daher nehmen wir den ATT als das für diese Studie bedeutsamere Maß an, obwohl keine allgemeinen Regeln besagen, wann die Verwendung von ATT oder ATE indiziert ist.

In PIRLS besteht die Zielgruppe aus Schüler/-innen der vierten Jahrgangsstufen der teilnehmenden Länder (Martin & Mullis, 2012). Hinsichtlich der Stichprobe und Population kann zwischen dem ATE und ATT der *Stichprobe* (SATE und SATT) und dem ATE und ATT der *Population* (PATE und PATT) unterschieden werden (vgl. DuGoff, Schuler & Stuart, 2014). Der SATE und SATT entsprechen ATE und ATT in ungewichteten Studienstichproben, welche ausschließlich auf Schüler/-innen der Stichproben generalisiert werden können. Im Gegensatz dazu nutzen PATE und PATT gewichtete Stichproben und beziehen das Studiendesign mit ein, sodass für die Zielpopulation generalisierte Aussagen getroffen werden können. PATE und PATT finden, wie auch im vorliegenden Fall, häufigere Anwendung, da die Generalisierbarkeit auf Populationen für viele Fragestellungen von zentraler Bedeutung ist.

2.2 Der Propensity Score

In Experimentalstudien werden Selektionsmechanismen durch die randomisierte Zuweisung von Individuen zu Experimental- und Kontrollgruppe vermieden. Randomisierung gewährleistet, dass alle Kovariaten zwischen den Gruppen balanciert sind, sodass bei ausreichender Stichprobengröße – abgesehen von dem Treatment – keine Unterschiede bestehen. Propensity Score Methoden imitieren randomisierte Studien, indem sie die Kovariaten post hoc balancieren. Dass eine solche Adjustierung mittels Propensity Score die Verzerrung der beobachteten konfundierenden Variablen entfernt, gehört zu den Grundannahmen von Rosenbaum und Rubin (1983). Der Propensity Score e_i wird durch die bedingte Wahrscheinlichkeit der Treatmentgruppenzugehörigkeit definiert, konditioniert auf eine Reihe von Kovariaten:

$$e_i = \Pr(A_i = 1 \mid X_i)$$

Mithilfe von logistischer Regression wird der Propensity Score, also die vorhergesagte Wahrscheinlichkeit, dass ein Kind das Treatment erhält, für jedes Individuum berechnet. Im vorliegenden Fall stellen der Indikator für das Nichtbesuchen von Kitas die abhängige Variable und die Kovariaten die unabhängigen Variablen dar. Bedingt für diese Werte weisen die Kovariaten X_i in den zwei Gruppen, der Kontroll- und Treatmentgruppe, die gleichen Verteilungen auf. Mit dem Propensity Score wird das Problem des Selektionsmechanismus' aus einem multivariaten Vektor in einen univariaten Score e_i für jedes Individuum übertragen. Hierbei bleibt zu beachten, dass auf diese Weise zwar die beobachteten Kovariaten ausbalanciert werden können, im Gegensatz zur Randomisierung jedoch nicht die unbeobachteten. Die Grundannahme bei der Anwendung von Propensity Score Methoden ist demnach, dass keine weiteren Verzerrungen durch unbeobachtete Kovariaten vorhanden sind (Rosenbaum & Rubin, 1983; Imbens, 2004). Stuart (2010) argumentiert, dass diese Annahme häufig gerechtfertigt ist, da durch die Kontrolle der Kovariaten (zumindest teilweise) auch unbeobachtete Kovariaten mitkontrolliert werden, sofern diese miteinander korrelieren. Die Achillesferse von Propensity Score Methoden stellen demnach nur solche unbeobachteten Kovariaten dar, die *nicht* mit beobachteten Kovariaten korrelieren. Anzumerken ist, dass Propensity Score Methoden für (nicht randomisierte) Beobachtungsstudien mit multiplen Treatments (Imai & Van Dyk, 2004; McCaffrey et al., 2013), Ausprägungsintensitäten des Treatments (Imbens, 2000; Rosenbaum, 2002) und Treatments auf Schulebene (Stuart, 2007) vorliegen.

2.3 Propensity Score Matching

Die primären Herangehensweisen an die Adjustierung der Gruppenzugehörigkeit mittels Propensity Score sind Propensity Score Matching, Stratifikation und Gewichtung (Austin, 2011; Stuart, 2010). In der vorliegenden Studie wird das Verfahren des Matchings angewendet. Die einfachste Form ist die Methode des 1:1 Nearest Neighbor Matchings. Hierbei wird jeweils eine Beobachtung aus der Treatmentgruppe mit derjenigen Beobachtung aus der Kontrollgruppe gematcht, welche den ähnlichsten Propensity Score aufweist. Insgesamt kann eine Vielzahl an Matching-Methoden genutzt werden, die sich durch verschiedene Spezifikationen unterscheiden: Matching mit oder ohne Zurücklegen, das Verhältnis k:1 Fälle im Vergleich von Kontroll- und Treatmentgruppe oder Caliper-Matching, bei dem vorher festgelegt wird, um wie viel die gematchten Propensity Scores maximal voneinander abweichen dürfen. Auf Basis des Nearest Neighbor-Verfahrens können sowohl SATT als auch PATT berechnet werden, da jede Beobachtung der Treatmentgruppe mit Beobachtungen aus der Kontrollgruppe gematcht und der Rest der Kontrollgruppe verworfen wird. In diesem Fall werden nicht alle verfügbaren Daten verwendet, da Kontrollgruppenbeobachtungen ausgeschlossen werden, sobald andere einen besser passenden Propensity Score aufweisen.

2.4 Maße zur Beurteilung der Balancierung

In querschnittlichen Beobachtungsstudien können so deutliche Unterschiede zwischen der Kontroll- und Treatmentgruppe bestehen, dass diese durch das Anwenden von Propensity Score Methoden nicht effektiv ausgeglichen werden können. Daher stellt die diagnostische Betrachtung der Balancierung der Kovariaten nach dem Matching einen zentralen Bestandteil des Verfahrens dar. Rubin (2001) schlägt die Beurteilung dreier Kriterien vor: der Standardized Bias, das Verhältnis der Varianzen der Propensity Scores und das Verhältnis der Residuenvarianzen der Kovariaten nach der Regression auf den Propensity Score. Der Standardized Bias wird als der Gruppenmittelwertunterschied dividiert durch die Standardabweichung der Treatmentgruppe definiert. Dieses Maß kann für binäre und kontinuierliche Variablen berechnet werden. Idealerweise sollte dieser Wert so klein wie möglich sein, jedoch werden Werte unter 0.25 als akzeptabel angesehen (z. B. Harder, Stuart & Antony, 2010; Stuart, 2010). Da diese Konvention eher einer Faustregel gleichkommt, sollten verbleibende Unterschiede durch Regression adjustiert werden (s. folgender Absatz; Austin, 2009). Das Verhältnis der Varianzen der Propensity Scores und das Verhältnis der Re-

siduen der Kovariaten nach der Adjustierung sollten jeweils nahe 1 liegen (beispielsweise zwischen 0.5 und 2; Rubin, 2001).

2.5 Schätzung des kausalen Effekts

Propensity Score Matching stellt per se keine Methode dar, um kausale Effekte zu schätzen, sondern um möglichst ähnliche Kontroll- und Treatmentgruppen zu bilden. Die Schätzung der kausalen Effekte folgt in einem separaten Schritt nach der Adjustierung. Hierzu können sowohl parametrische als auch nicht-parametrische Ansätze genutzt werden. Nach dem Propensity Score Matching kann die abhängige Variable der gematchten Stichprobe beispielsweise auf die binäre Variable der Gruppenzugehörigkeit regressiert werden, um deren kausalen Effekt zu schätzen. Ein Vorteil der Anwendung von Regressionsmodellen ist, dass die beobachteten Kovariaten als Kontrollvariablen in das Modell mit aufgenommen werden können. Solche über das Matching hinausgehenden Adjustierungen kontrollieren verbleibende Unterschiede zwischen den Gruppen weitergehend, falls keine perfekte Balancierung erreicht werden konnte. Um den PATT zu berechnen, kann eine gewichtete Regression mit den studienspezifischen Gewichten angewendet werden, wodurch den spezifischen Eigenschaften von internationalen Vergleichsstudien Rechnung getragen werden kann (Verwendung von Plausible Values, komplexe Stichprobendesigns, etc.). In diesem Fall müssen alle Analysen für die verschiedenen Plausible Values und unter Einbezug von Replikationstechniken für die Schätzung der Varianzen durchgeführt werden (beispielsweise Jackknife 2 für PIRLS; s. Foy & Kennedy, 2008).

3 Analysen: Die Anwendung von Propensity Score Matching auf PIRLS Daten

3.1 Stichprobe

Um die Frage nach dem Effekt des fehlenden Besuchs einer Kita auf spätere Lesekompetenzen zu untersuchen, wenden wir Propensity Score Matching Methoden auf Daten aus PIRLS 2001, 2006 und 2011 in Deutschland an (Mullis, Martin, Gonzalez & Kennedy, 2003; Mullis, Martin, Kennedy & Foy, 2007; Mullis et al., 2012). PIRLS beinhaltet neben Lesekompetenzmaßen auch Informationen über die Teilnahme an Kitas und Faktoren des familiären und weiteren Lernumfelds. In der vorliegenden Studie nutzen wir Informationen aus den Lesekompetenztests sowie den Fragebögen der

Schüler/-innen und Eltern, die Trendvergleiche über die Erhebungszyklen hinweg zulassen. Die Analysen basieren auf stratifizierten und geclusterten Zufallsstichproben von Schüler/-innen der vierten Jahrgangsstufe aus drei verschiedenen Erhebungszyklen ($n_{2001} = 7633$; $n_{2006} = 7899$; $n_{2011} = 4000$).

3.2 Instrumente

3.2.1 Treatment

Um den interessierenden Effekt möglichst illustrativ betrachten zu können, verzichten wir darauf, die Kitateilnahme kontinuierlich zu operationalisieren und eine abgestufte Ausprägung der Teilnahme zu betrachten (vgl. Imai & van Dyk, 2004). Vielmehr verwenden wir eine binäre Treatmentvariable, bei der 0 dafür steht, dass ein Kind mehr als zwei Jahre lang eine Kita besucht hat, und der Wert 1 indiziert, dass es keine Kita besucht hat. Wir untersuchen also kontrastierend Effekte des Treatments ‚keine Kitateilnahme' versus die Kontrolle ‚intensive Kitateilnahme' und schließen Kinder aus, die eine Kindertageseinrichtungen maximal zwei Jahre lang besucht haben.

Auf Basis des Propensity Score Matchings können wir den ATT berechnen; wir schätzen also den Effekt, den die Kitateilnahme für solche Kinder hätte, die keine Kita besucht haben. Diesen Effekt nehmen wir als am interessantesten für politische Entscheidungen über weiter zu steigernde Kitateilnahmequoten an. Die Informationen zur Kitateilnahme entstammen zwei Items aus den Elternfragebögen: Die Items ASBH04A und ASBH04B erfassen, ob das Kind überhaupt eine Kita besucht hat und wie lange es vor der Schule eine Kita besucht hat, falls es eine besucht hat (Itembezeichnungen entstammen den Originaldatensätzen aus PIRLS 2011). Basierend auf diesen Items konnten somit die Kinder identifiziert werden, die Kitas nicht beziehungsweise mehr als zwei Jahre besucht haben. Im Jahr 2001 ging aus den Elternfragebögen hervor, dass 293 Kinder keine Kita besucht hatten, 196 in 2006 und 56 in 2011. Diese bilden die Treatmentgruppen. Demgegenüber besuchten 5321 Kinder im Jahr 2001 mehr als zwei Jahre lang eine Kita, 6674 in 2006 und 3612 in 2011. Diese bilden die Kontrollgruppengrundlagen für das Propensity Score Matching.

3.2.2 Outcome

Als abhängige Variable ziehen wir die Lesekompetenzgesamtpunktzahl heran (ASRREA01-ASRREA05; weitere technische Ausführungen in von Davier, Gonzalez & Mislevy, 2009; Martin & Mullis, 2012). Alle fünf Plausible Values werden in die Analysen einbezogen und mittels Rubins (1987)

Regeln verbunden. In Tabelle 1 sind deskriptive Statistiken des Outcomes dargestellt.

3.2.3 Kovariaten

Die Verfügbarkeit und Auswahl von beobachteten Variablen, die als Kovariaten herangezogen werden, beeinflussen die interne Validität des Matchings maßgeblich. Um zu modellieren, welche Kovariaten Einfluss auf die Selektionsmechanismen des Kitabesuchs nehmen, folgen wir dem Personen-Prozess-Kontext-Zeit-Modell und berücksichtigen den sozio-ökonomischen Status und weitere Hintergrundmerkmale der Kinder und Familien (vgl. Abschnitt 1). Internationale empirische Evidenz zeigt, dass solche Faktoren länderübergreifend zentrale Kovariaten der Teilnahme und Nichtteilnahme an Kitas darstellen (vgl. Grogan, 2012; Hirshberg, Huang & Fuller, 2005; Kim & Fram, 2009; Müller, Strietholt & Hogrebe, 2014; Zachrisson, Janson & Nærde, 2013; Vandenbroeck, De Visscher, Van Nuffel & Ferla, 2008). Daher kombinieren wir folgende Hintergrundvariablen aus den Fragebögen für Schüler/-innen und Eltern, die den Teilnahmestatus an Kitas oder die Lesekompetenz beeinflussen (s. Tab. 1 für deskriptive Statistiken):

- Das *Geschlecht* (ITSEX) der Schüler/-innen wurde bereits in der Phase der Stichprobenziehung erfasst.
- Die *zuhause gesprochene Sprache* (ASBG03) wird mit einem Einzelitem aus dem Fragebogen für Schüler/-innen erfasst, das abfragt, wie häufig zuhause Deutsch gesprochen wird. Das Item weist ein dreistufiges Likert-Antwortformat auf („Ich spreche immer Deutsch", „Ich spreche manchmal Deutsch und manchmal eine andere Sprache", „Ich spreche niemals Deutsch").
- Die *Anzahl zuhause vorhandener Bücher* wurde sowohl von den Schüler/-innen (ASBG04) auf einer fünfstufigen Likert-Skala („Keine oder nur sehr wenige (0–10 Bücher)", „Genug, um ein Regalbrett zu füllen (11–25 Bücher); „Genug, um ein Regal zu füllen (26–100 Bücher)", „Genug, um zwei Regale zu füllen (101–200 Bücher)", „Genug, um drei oder mehr Regale zu füllen (über 200 Bücher)") als auch Eltern eingeschätzt (ASBH14; „0–10", „11–25", „26–100", „101–200", „Über 200"). Beide Variablen werden separat aufgenommen, da die Einschätzungen zwischen Kindern und Eltern variieren können (z. B. Jerrim & Micklewright, 2014).

- Die *elterliche Einstellung zum Lesen* ist ein eigens berechneter Index, der die mittlere Ausprägung der vier Items ASBH13A, ASBH13C, ASBH13D und ASBH13E aus dem Elternfragebogen abbildet, welche erfragen, wie gern die Eltern selbst lesen. Die Items liegen jeweils auf einer vierstufigen Likert-Skala („Ich stimme stark zu", „Stimme einigermaßen zu", „Stimme wenig zu", „Stimme überhaupt nicht zu").
- Die *höchste elterliche berufliche Beschäftigung* (ASDHOCCP) wird aus zwei Elternfragebogenitems abgeleitet, in denen siebenstufig nach dem Hauptberuf von Vater und Mutter gefragt wird (z. B. „Hat niemals bezahlte Arbeit außerhalb des Haushalts verrichtet").
- Der *höchste elterliche Schulabschluss* (ASDHEDUP) ist eine Variable, die aus zwei Elternfragebogenitems gebildet wird, in denen der höchste erreichte Schulabschluss von Vater und Mutter auf der fünfstufigen ISCED-Skala abgefragt wird.
- *Frühe häusliche Bildungsaktivitäten* werden als ein eigens berechneter Index aufgenommen, der die mittlere Ausprägung der sechs Items ASBH02A, ASBH02B, ASBH02C, ASBH03D, ASBH03G und ASBH02I des Elternfragebogens widerspiegelt. Diese beziehen sich auf die Häufigkeit von einer Bandbreite an bildungsförderlichen Aktivitäten im frühen Kindesalter. Die Items liegen jeweils auf einer dreistufigen Likert-Skala („Oft", „Manchmal", „Nie oder fast nie").

Aus Gründen der Einfachheit betrachten wir alle Kovariaten abgesehen vom Geschlecht als kontinuierlich. Es wurden ausschließlich solche Variablen als Kovariaten aufgenommen oder als Index zusammengefasst, die zu allen drei Messzeitpunkten sowohl in Hinblick auf die Itemformulierung als auch die Antwortformate identisch erfasst wurden. Eine generelle Limitation der genannten Kovariaten ist, dass diese nicht vor der Kitateilnahme beziehungsweise Nichtteilnahme erhoben wurden, sondern erst am Ende der Grundschulzeit. Daher könnten diese nicht nur Einfluss auf die Selektionsmechanismen der Teilnahme genommen haben, sondern auch andersherum. Dies hätte ein massives Bedrohungspotenzial für die vorliegende Studie zur Folge, da wir auf ein Outcome der Kitateilnahme konditionieren würden. Beispielsweise wird die Verfügbarkeit öffentlich geförderter Frühbetreuung auch als Maßnahme angesehen, die Berufstätigkeit insbesondere von Müttern zu befördern. Wenn solche Kovariaten selbst auch Outcomes sind, dann sind sie ungeeignete Kontrollvariablen und sollten nicht aufge-

nommen werden, da dies in einer Unterschätzung der Treatmenteffekte resultieren würde (z. B. Angrist & Pischke, 2009). Im vorliegenden Fall gehen wir davon aus, dass es sich um zeitlich stabile Merkmale handelt, da niedrigschwellige Kitaangebote beispielsweise vornehmlich die Berufstätigkeit eines Elternteils beeinflussen sollten. Da wir die *höchste elterliche berufliche Beschäftigung* betrachten, also eine Kombination der Berufstätigkeit beider Eltern, gehen wir nicht davon aus, dass Kitaangebote diese deutlich beeinflussen.

3.3 Umgang mit fehlenden Werten

Fehlende Werte wurden mittels Predictive Mean Matching fünfmal imputiert (z. B. Rubin, 1987). Die Imputationen wurden für die drei Messzeitpunkte separat und unter Einschluss der genannten Treatment- und Outcomevariablen sowie Kovariaten durchgeführt. Im Fall der Outcomes wurde lediglich ein Plausible Value in die Imputationen eingeschlossen. Die Anteile an fehlenden Werten reichten von 4.4% bis 19.1% bei den Variablen aus den Fragebögen für Schüler/-innen und von 13.3% bis 43.6% bei den Elternfragebogenmaßen (vgl. Tab. 1). Im Fall des Outcomes lagen keine fehlenden Werte vor. Die imputierten Datensätze wurden jeweils mit einem der fünf Plausible Values der Lesekompetenz verknüpft. Alle folgenden Analysen wurden für jeden imputierten Datensatz durchgeführt und mittels Rubins (1987) Regeln kombiniert.

Tabelle 1: Deskriptive Statistiken der Originalstichproben

	2001 n = 7633			2006 n = 7899			2011 n = 4000		
	M	SSD	% fehlend	M	SSD	% fehlend	M	SSD	% fehlend
Leseleistung	534.77	65.13	0.00	547.57	61.25	0.00	541.56	62.02	0.00
Geschlecht	1.51	0.50	0.04	1.51	0.50	0.00	1.51	0.50	0.00
Zuhause gesprochene Sprache	1.88	0.35	4.40	1.71	0.47	19.10	1.79	0.43	10.10
Anzahl Bücher (Eltern)	2.47	1.20	13.32	2.61	1.19	13.61	2.55	1.22	20.73
Anzahl Bücher (Schüler/ innen)	2.03	1.18	7.05	2.14	1.20	10.24	2.16	1.16	10.75
Elterliche Einstellung zum Lesen	2.11	0.78	17.92	2.18	0.74	18.76	2.12	0.77	23.75
Höchste elterliche berufliche Stellung	3.65	1.10	43.60	3.52	1.14	21.27	3.58	1.16	27.18
Höchster elterlicher Schulabschluss	2.74	1.06	35.92	2.12	1.13	25.17	2.47	1.26	27.05
Fruhe häusliche Bildungsaktivitäten	1.17	0.39	17.65	1.29	0.38	17.67	1.35	0.38	22.77

3.4 Modellierung

Die Stichproben aus 2001, 2006 und 2011 umfassen zwischen 56 und 293 Kinder, die keine Kita besucht haben. Jedes dieser Kinder wird mit einem Kind gematcht, das aus derselben Kohorte stammt, mindestens zwei Jahre eine Kita besucht hat und einen ähnlichen Propensity Score aufweist. Dementsprechend wird die 1:1 Nearest Neighbor Matching Methode ohne Zurücklegen angewendet, sodass jedes Kind aus der Kontrollgruppe nur einem Kind aus der Treatmentgruppe zugeordnet wird. Das Matching impliziert die Schätzung des ATT für Kinder, die keine Kita besucht haben, d. h. die Schätzung der Langzeitkonsequenzen der Nichtteilnahme für die Lesekompetenz in der vierten Jahrgangsstufe. Für die Berechnung individueller Propensity Scores werden logistische Regressionsmodelle verwendet, in denen das binäre Treatment auf die Kovariaten regressiert wird. Das Matching wird sowohl für die drei Messzeitpunkte als auch die fünf imputierten Datensätze separat durchgeführt, woraus effektive gematchte Stichprobengrößen von $n_{2001} = 587$, $n_{2006} = 312$ und $n_{2011} = 112$ resultieren. Die ungerade Stichprobengröße in 2001 ergibt sich aus der Imputationsvarianz bei der Treatmentvariable. Im Anschluss wird die Balancierung dieser gematchten Stichproben anhand des Standardized Bias, des Verhältnisses der Varianzen der Propensity Scores und des Verhältnisses der Residuenvarianzen der Kovariaten nach der Regression auf den Propensity Score geprüft.

In dem Analysemodell zur Beantwortung der Forschungsfrage nach dem Effekt der Nichtteilnahme an Kitas auf spätere Lesekompetenzen regressieren wir das Outcome auf die Treatmentvariable für jede gematchte Stichprobe. Da die Treatmentvariable binär ist, schätzen wir die Mittelwertdifferenz der Lesekompetenz zwischen Treatment- und Kontrollgruppe. In weiteren Analysen ergänzen wir diese Regression um die Matching-Kovariaten, um kleinere verbleibende Gruppenunterschiede in der Balancierung des Matchings zu adjustieren.

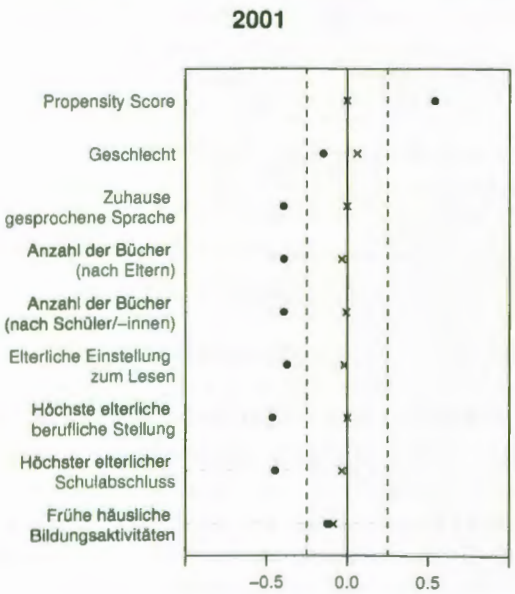
Um die Befunde auf die Gesamtpopulationen von Viertklässler/-innen in Deutschland in 2001, 2006 und 2011 zu generalisieren, wird der PATT geschätzt. Daher wird das PIRLS Stichprobengewicht HOUWGT in allen Regressionsmodellen berücksichtigt. Darüber hinaus wenden wir die Jackknife Repeated Replication Technik unter Verwendung der Variablen JKZONE und JKREP an, um bei der Schätzung der Standardfehler den stratifizierten geclusterten Stichproben in PIRLS angemessen Rechnung zu tragen (Lohr, 2010; Rutkowski, Gonzalez, Joncas & von Davier, 2010). Alle Analysen werden in der R Umgebung durchgeführt. Für die Imputation fehlender Werte nutzen wir das Paket *MICE* (Buuren & Groothuis-Oudshoorn, 2011), für das Matching *MatchIt* (Ho, Imai, King & Stuart, 2011)

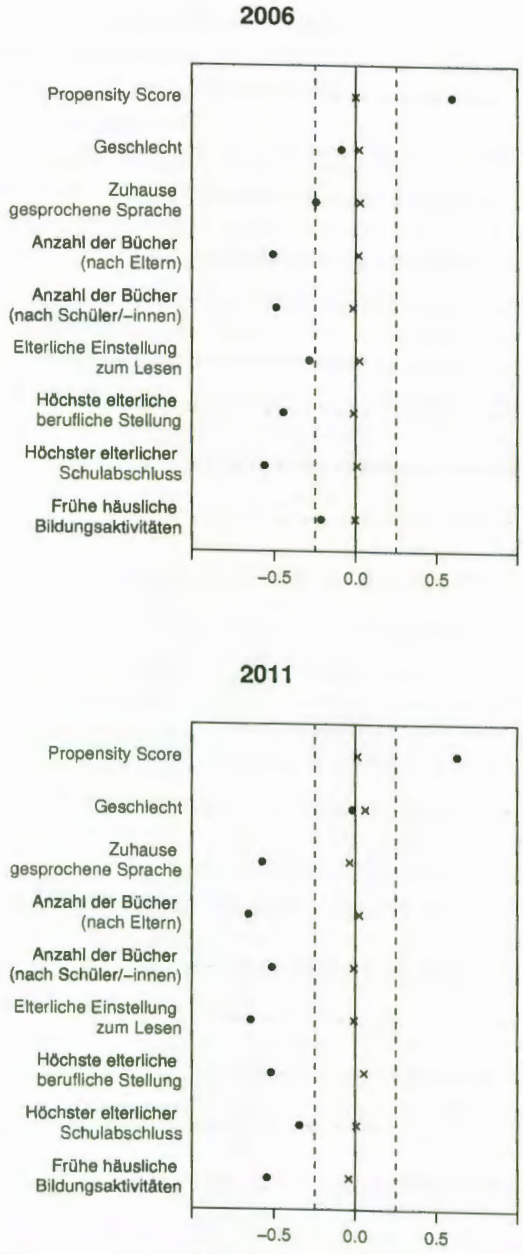
und *Survey* (Lumley, 2014) für Regressionsanalysen mit der Jackknife Repeated Replication Technik.

4 Ergebnisse

Vor dem Matching lässt sich in Bezug auf die meisten Kovariaten feststellen, dass Kinder, die mehr als zwei Jahre lang eine Kita besucht haben, häufiger in leselernförderlichen Haushalten aufgewachsen sind, als Kinder, die keine Kita besucht haben. Zum Beispiel zeigt sich in Bezug auf die *höchste elterliche berufliche Stellung* zu allen Messzeitpunkten ein Unterschied von circa einer halben Standardabweichung zu Ungunsten der Kinder ohne Kitateilnahme. In Abbildung 1 ist für jede Kovariate sowohl der Standardized Bias vor dem Matching (Punkte) als auch nach dem Matching (Kreuze) abgetragen. Negative Werte bedeuten, dass Kinder, die keine Kita besucht haben, beispielweise weniger Bücher zuhause haben oder weniger häufig Deutsch in der Familie sprechen.

Abbildung 1: Standardized Bias in ungematchten und gematchten Stichproben





Anmerkung. In den Abbildungen indiziert • die standardisierte Mittelwertdifferenz zwischen Kontroll- und Treatmentgruppe vor dem Matching und × nach dem Matching (gestrichelte Linie -0.25/+0.25).

Der Standardized Bias ist die Mittelwertdifferenz zwischen Treatment- und Kontrollgruppe geteilt durch die Standardabweichung der Treatmentgruppe. Die Treatmentgruppe wird herangezogen, um die Mittelwertdifferenzen zu standardisieren, da diese vor und nach dem Matching identisch zusammengesetzt ist, während die Kontrollgruppe durch das Matching reduziert wird. Nach dem Matching lagen die standardisierten Mittelwertdifferenzen aller Kovariaten zu jedem Messzeitpunkt unter 0.25 (siehe den Bereich zwischen den gestrichelten Linien in Abb. 1). Dies indiziert einen geringfügigen Bias nach dem Matching. Auch die Vergleiche der Varianzen der Propensity Scores und der Verhältnisse der Residuenvarianzen nach der Propensity Score Adjustierung liefern weitere Belege dafür, dass das Matching gut vergleichbare Treatment- und Kontrollgruppen erzielen konnte. Die Varianzverhältnisse der Propensity Scores liegen zwischen 0.86 und 1.01 und die der Residuen der Kovariaten zwischen 0.73 und 1.18, also jeweils nahe 1 (vgl. Rubin, 2001).

4.1 Gruppenunterschiede vor dem Matching

Die Koeffizienten einer Reihe von Regressionsanalysen sind in Tabelle 2 verzeichnet. In dem oberen Teil werden die Ergebnisse der Modelle dargestellt, die auf den ungematchten Stichproben aus 2001, 2006 und 2011 der Kinder basieren, die keine Kita besucht haben, und den Kindern, die mehr als zwei Jahre eine Kita besucht haben. Die Konstante ist die mittlere Lesekompetenz solcher Schüler/-innen, die mehr als zwei Jahre an Kitabetreuung teilgenommen haben, während der Treatmentschätzer die mittlere Abweichung der Nichtteilnehmenden repräsentiert.

Es zeigt sich, dass die Kinder, die mehr als zwei Jahre an Kitabetreuung teilgenommen haben, signifikant bessere Lesekompetenzen aufweisen, als die Gleichaltrigen ohne Kitabesuch. In 2001 und 2006 beträgt der Unterschied circa 29 Punkte auf der PIRLS Leseskala und in 2011 sogar circa 40 Punkte. Diese beobachteten Leistungsunterschiede zwischen Treatment- und Kontrollgruppe sind in den ungematchten Stichproben mit Kovariaten der Kitateilnahme konfundiert.

4.2 Gruppenunterschiede nach dem Matching

Als Folge des Matchings gleichen sich die Koeffizienten der Kontrollgruppe an die der Treatmentgruppe an, da nun nur solche Kinder verglichen werden, die abgesehen von der Kitabetreuung sehr ähnlichen Hintergrundeinflüssen ausgesetzt waren. Daher kann der PATT auf Basis der gematchten Stichprobe geschätzt werden, sobald Stichprobengewichte berücksichtigt

werden. Über die drei Messzeitpunkte hinweg zeigt sich in den Regressionsanalysen, dass Schüler/-innen, die keine Kita besucht haben, geringfügig niedrigere Lesekompetenzen aufweisen, als solche, die mehr als zwei Jahre in Kitas betreut wurden. Gleichzeitig sind diese Differenzen zu keinem Messzeitpunkt statistisch signifikant auf dem 5%-Niveau (vgl. Mittelteil der Tab. 2).

In weiteren Analysen wurden die Matchingvariablen als Kontrollen in die Regressionsmodelle miteingeschlossen, um verbleibende Gruppenunterschiede in den Kovariaten zu adjustieren. Die resultierenden Koeffizientenschätzungen weichen in allen drei PIRLS-Zyklen kaum von den Ergebnissen der Regressionen ohne Kontrollen ab (vgl. unterer Teil der Tab. 2). Auch hier sind leichte Leseleistungsnachteile solcher Schüler/-innen, die keine Kita besucht haben, nicht signifikant von null verschieden. Dieser Befund stützt weiterhin, dass durch das Matching eine angemessene Balancierung zwischen Treatment- und Kontrollgruppe erzielt werden konnte.

Tabelle 2: Regression der Leseleistung auf das Nichtbesuchen von Kitas

	2001	2006	2011
Ungematchte Stichprobe			
Konstante	544.91** (1.89)	550.32** (2.15)	543.23** (2.33)
Treatment	-28.67** (8.17)	-28.73* (13.28)	-40.34** (10.34)
Kovariaten	nein	nein	nein
N	5614	6967	3668
Gematchte Stichprobe			
Konstante	525.14** (8.94)	531.52** (11.98)	516.46** (10.11)
Treatment	-8.90 (8.79)	-9.93 (15.21)	-13.57 (14.63)
Kovariaten	nein	nein	nein
N	587	392	112
Gematchte Stichprobe mit Kovariaten als Kontrollen			
Konstante	392.73** (18.94)	445.71** (37.07)	433.34** (45.69)
Treatment	-9.96 (8.23)	-12.51 (13.22)	-11.73 (11.38)
Kovariaten	ja[a]	ja[a]	ja[a]
N	587	392	112

Anmerkung. Die abhängige Variable ist die Lesekompetenzgesamtpunktzahl. * = $p < 0.05$, ** = < 0.01.
[a] Als Kontrollvariablen aufgenommen wurden: *Geschlecht, zuhause gesprochene Sprache, Anzahl Bücher (Schüler/-innen), Anzahl Bücher (Eltern), elterliche Einstellung zum Lesen, höchste elterliche berufliche Beschäftigung, höchster elterlicher Schulabschluss* und *frühe häusliche Bildungsaktivitäten.*

5 Diskussion und Schlussfolgerung

Kinder, die aus bildungsbenachteiligten familiären Kontexten stammen, besuchen seltener Kitas. Auch wenn die Teilnahmequoten laut den PIRLS Elternbefragungen über die Erhebungszyklen hinweg gestiegen sind, zeigen

sich über die Jahre vergleichbare Selektionsmechanismen. Um den kausalen Effekt der Nichtteilnahme an Kitas auf die spätere Leseleistung zu bestimmen, wurden teilnehmende und nicht-teilnehmende Kinder erfolgreich gematcht. Weder in 2001, 2006 noch 2011 konnte ein signifikanter Unterschied in der Lesekompetenz der Treatment- und Kontrollgruppen gefunden werden. Die vorliegende Studie zeigt damit keine empirische Evidenz für die Hypothese, dass die Teilnahme an Kindertagesbetreuung bei Kindern, die keine Kita besucht haben, zu verbesserten Lesekompetenzen in der vierten Klasse führen würde. Zu allen Messzeitpunkten wurde jedoch ein geringfügiger Lesekompetenznachteil solcher Kinder gefunden, die keine Kita besucht hatten. Dieser Befund blieb bestehen, sobald in weiteren Analysen die Kovariaten des Matchings als Kontrollvariablen in die Regressionsanalysen aufgenommen wurden.

Die Studienergebnisse stehen im Einklang mit den Erkenntnissen von Hogrebe und Strietholt (2016), die nur für Schweden und Singapur, nicht jedoch für Deutschland gezeigt haben, dass die Teilnahme an institutioneller frühkindlicher Bildung positive Auswirkungen auf bildungsbenachteiligte Kinder ohne Kitabesuch gehabt hätte. Dieser Befund stimmt auch mit anderen Ergebnissen vorheriger Studien überein, in denen nur Leistungseffekte für weniger bildungsbenachteiligte Kinder gefunden wurden (vgl. Vogel et al., 2010). Da eine hohe Programmqualität international als Kernvoraussetzung für die erfolgreiche Umsetzung frühkindlicher Förderung identifiziert werden konnte, lassen sich die Befunde der vorliegenden Studie so interpretieren, dass die Kitaeinrichtungen in Deutschland nach wie vor nicht so ausgestaltet sind, dass sie bildungsbenachteiligte Kinder hinreichend unterstützen können (s. auch Duncan & Magnuson, 2013). Die Tatsache, dass über die Erhebungszyklen hinweg leicht voneinander verschiedene ATTs gefunden wurden, könnte jedoch darauf hindeuten, dass sich die Selektionsmechanismen oder Qualitätsmerkmale von Kitas mit den Jahren verändern.

Gleichzeitig müssen die Ergebnisse vor dem Hintergrund der Limitationen der Studie interpretiert werden. Propensity Score Methoden können verwendet werden, um ein randomisiertes Experiment auf Basis querschnittlicher Beobachtungsdaten zu imitieren. Gleichzeitig unterliegt dieses Vorgehen der Annahme, dass die Zuordnung zum Treatment nicht von weiteren, unbeobachteten Kovariaten abhängt, die wiederum mit dem Outcome korrelieren (Rosenbaum & Rubin, 1983). Obwohl in PIRLS eine Vielzahl an Hintergrundvariablen verfügbar ist, könnten die vorliegenden Variablen den familiären Hintergrund nicht perfekt messen oder alle relevanten Facetten abdecken. Die Reliabilität der Informationen über frühe häusliche

Bildungsaktivitäten könnte beispielsweise limitiert sein, da sie erst in der Grundschulzeit des Kindes erfasst wurden. Aufgrund unbeobachteter Kovariaten könnten also auch nach dem Matching systematische Gruppenunterschiede zwischen Teilnehmenden und Nichtteilnehmenden bestehen bleiben. Obwohl es unmöglich ist, solche Einflüsse von Kovariaten vollständig vorherzusagen oder auszuschließen, scheint es lohnenswert, über deren hypothetische Auswirkungen nachzudenken: Falls Schüler/-innen, die keine Kita besucht haben, auch in Bezug auf weitere Prädiktoren der Lesekompetenz benachteiligt wären, würden die Gruppendifferenzen überschätzt, die im vorliegenden Fall bereits nicht signifikant sind. Es scheint demnach nicht naheliegend, dass die weitere Berücksichtigung von Hintergrundmerkmalen Ergebnisse generieren würde, die der Kitateilnahme einen signifikanten Effekt in Bezug auf die Lesekompetenzen zuschreiben würde. Obwohl unsere Analysen robustere empirische Befunde liefern können als korrelative Untersuchungen, schränkt die fehlende randomisierte Zuweisung zur Kitateilnahme dennoch die Möglichkeit ein, kausale Schlussfolgerungen zu ziehen.

In unserer Studie schätzten wir den Effekt der Kitateilnahme für solche Kinder, die tatsächlich keine Kita besucht haben (PATT). Diese Kinder kommen durchschnittlich aus bildungsbenachteiligten Elternhäusern, sodass die Befunde lediglich auf solche und keinesfalls auf alle Kinder in Deutschland übertragen werden sollten. Darüber hinaus ist deutlich, dass die gefundene Ineffektivität für die Lesekompetenz in der vierten Jahrgangsstufe nicht bedeutet, dass keine förderlichen Effekte für andere schulische Leistungen oder schulische Leistungen zu früheren oder späteren Zeitpunkten der Schullaufbahn bestehen (zu nachlassenden Effekten vgl. Barnett, 2011; Duncan & Magnuson, 2013).

Gleichwohl können unsere Befunde als bedeutsam für politische Entscheidungen angesehen werden, da wir weder für 2001, 2006 noch 2011 förderliche Effekte der Kitateilnahme für bildungsbenachteiligte Kinder in Bezug auf ihre Leseleistungen in der vierten Klasse gefunden haben. Unsere Studie stützt die Annahme, dass das niedrige Leistungsniveau dieser Kinder nicht aufgrund der Nichtteilnahme an Kindertagesbetreuung zustande kommt. Aus politischer Perspektive bedeuten diese Befunde, dass alternative Interventionen oder alternative Ansätze der Kindertagesbetreuung notwendig sind, um diese Kinder besser zu fördern.

Referenzen

Anders, Y. (2013). Stichwort: Auswirkungen frühkindlicher institutioneller Betreuung und Bildung. *Zeitschrift für Erziehungswissenschaft, 16*(2), 237–275. doi: 10.1007/s11618-013-0357-5

Anders, Y. & Roßbach, H.-G. (2014). Empirische Bildungsforschung zu Auswirkungen frühkindlicher, institutioneller Bildung: Internationale und nationale Ergebnisse. In Braches-Chyrek, R., Rühner, Ch., Sünker, H. & Hopf, M. (Hrsg.), *Handbuch frühe Kindheit* (S. 335–347). Opladen: Barbara Budrick.

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*(25), 3083–3107. doi: 10.1002/sim.3697.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399–424. doi: 10.1080/00273171.2011.568786.

Barnett, W. S. (2011). Effectiveness of early education interventions. *Science, 333*(6045), 975–978. doi: 10.1126/science.1204534.

Bronfenbrenner, U. (1990). The ecology of cognitive development. *Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 10*(2), 101–114.

Bronfenbrenner, U. & Morris, P. A. (2006). The bioecological model of human development. In W. Damon & R. M. Lerner (Hrsg.), *Handbook of child psychology* (6. Aufl., Bd. 1, S. 793–828). New York, NY: Wiley.

Burger, K. (2010). How does early childhood care and education affect cognitive development? An international review of the effects of early interventions for children from different social backgrounds. *Early Childhood Research Quarterly, 25*(2), 140–165. doi: 10.1016/j.ecresq.2009.11.001.

Buuren, S. V. & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3).

Camilli, G., Vargas, S., Ryan, S. & Barnett, B. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112*(3), 579–620.

Campbell, F. A. & Ramey, C. T. (1994). Effects of early intervention on intellectual and academic achievement: A follow up study of children from low-income families. *Child Development, 65*(2), 684–698. doi: 10.1111/j.1467-8624.1994.tb00777.x.

Chambers, B., Cheung, A., Slavin, R. E., Smith, D. & Laurenzano, M. (2010). *Effective early childhood education programs: A systematic review*. Retrieved from the Best Evidence Encyclopedia website: http://www.bestevidence.org/word/early_child_ed_Sep_22_2010.pdf.

Davier, M. v., Gonzalez, E. J. & Mislevy, R. J. (2009). What are plausible values and why are they useful? In: IERI Monograph Series: *Issues and Methodologies in Large-Scale Assessments* (Bd. 2, S. 9–36). Hamburg/Princeton NJ: IEA-ETS Research Institute.

DESTATIS (2015). Betreuungsquoten der Kinder unter 6 Jahren in Kindertagesbetreuung. Retrieved from https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Soziales/Sozialleistungen/Kindertagesbetreuung/Tabellen/Tabellen_Betreuungsquote.html

DuGoff, E. E., Schuler, M. & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research, 49*(1), 284–303. doi: 10.1111/1475-6773.12090.

Duncan, G. J. & Magnusson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives, 27*(2), 109–132. doi: 10.1257/jep.27.2.109.

Early, D. M. & Burchinal, M. R. (2001). Early childhood care: Relations with family characteristics and preferred care characteristics. *Early Childhood Research Quarterly, 16*(4), 475–497. doi: 10.1016/S0885-2006(01)00120-X.

Foy, P. & Kennedy, A. M. (Hrsg.) (2008). *PIRLS 2006 user guide for the international database*. Chestnut Hill, MA: Boston College.

Grogan, K. E. (2012). Parents' choice of pre-kindergarten: the interaction of parent, child and contextual factors. *Early Child Development and Care, 182*(10), 1265–1287. doi: 10.1080/03004430.2011.608127.

Harder, V. S., Stuart, E. A. & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234–249. doi: 10.1037/a0019623.

Heckman, J. J. & Robb, R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Hrsg.), *Drawing inferences from self-selected samples* (S. 63–107). New York, NY: Springer.

Hirshberg, D., Huang, D. S.-C. & Fuller, B. (2005). Which low-income parents select childcare? Family demand and neighborhood organizations. *Children and Youth Services Review, 27*(10), 1119–1148. doi: 10.1016/j.childyouth.2004.12.029.

Ho, D., Imai, K., King, G. & Stuart, E. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*. Retrieved from http://gking.harvard.edu/matchit.

Hogrebe, N. (2014). *Bildungsfinanzierung und Bildungsgerechtigkeit*. Wiesbaden: Springer Fachmedien Wiesbaden.

Hogrebe, N. & Strietholt, R. (2016). Does non-participation in pre-school affect children's reading achievement? International evidence from propensity score analyses. *Large-Scale Assessment in Education, 4*(1), 1–22.

Hynes, K. & Habasevich-Brooks, T. (2008). The ups and downs of child care: Variations in child care quality and exposure across the early years. *Early Childhood Research Quarterly, 23*(4), 59–574. doi: 10.1016/j.ecresq.2008.09.001.

Imai, K. & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99*(467), 854–866. doi: 10.1198/016214504000001187.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706–710. doi: 10.1093/biomet/87.3.706.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86*(1), 4–29. doi: 10.3386/t0294.

Imbens, G. W. & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences. An introduction*. New York, NY: Cambridge University Press.

Jerrim, J. & Micklewright, J. (2014). Socio-economic gradients in children's cognitive skills: Are cross-country comparisons robust to who reports family background? *European Sociological Review, 30*(6), 766–781. doi: 10.1093/esr/jcu072.

Kim, J. & Fram, M. S. (2009). Profiles of choice: Parents' patterns of priority in child care decision-making. *Early Childhood Research Quarterly, 24*(1), 77–91. doi: 10.1016/j.ecresq.2008.10.001.

Knudsen, E. I., Heckman, J. J., Cameron, J. L. & Shonkoff, J. P. (2006). Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proceedings of the National Academy of Sciences, 27*, 10155–10162. doi: 10.1073/pnas.0600888103.

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Boston, MA: Brooke/Cole.

Lumley, T. (2014). *Survey: analysis of complex survey samples*. R package version, 3.30.

Martin, M. O. & Mullis, I. V. S. (Hrsg.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R. & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine, 32*(19), 3388–3414. doi: 10.1002/sim.5753.

Müller, N., Strietholt, R. & Hogrebe, N. (2014). Ungleiche Zugänge zum Kindergarten. In K. Drossel, R. Strietholt & W. Bos (Hrsg.), *Empirische Bildungsforschung und evidenzbasierte Reformen im Bildungswesen* (S. 33–46). Münster: Waxmann.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J. & Kennedy, A. M. (2003). *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary Schools.* Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M. & Foy, P. (2007). *IEA's Progress in International Reading Literacy Study in Primary School in 40 Countries.* Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P. & Drucker, K. T. (2012). *PIRLS 2011. International results in reading.* Retrieved from http://timssandpirls.bc.edu/pirls2011/international-results-pirls.html.

Organization for Economic Development and Cooperation (OECD). (2013). *PISA 2012 Results: Excellence through equity: Giving every student the chance to succeed* (Bd. II). Paris: OECD Publishing.

Pianta, R., Barnett, W., Burchinal, M. & Thornburg, K. (2009). The effects of preschool education: What we know, how public policy is or is not aligned with the evidence base, and what we need to know. *Psychological Science in the Public Interest, 10*(2), 49–88. doi: 10.1177/1529100610381908.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., Mashburn, A. & Downer, J., (2012). *Third grade follow-up to the Head Start impact study. Final report* (OPRE report no. 2012–45). Retrieved from the Administration for Children and Families, US Department of Health and Human Services website: http://www.acf.hhs.gov/sites/default/files/opre/head_start_report.pdf.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. doi: 10.1093/biomet/70.1.41.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. doi: 10.1037/h0037350.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York, NY: Wiley.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3–4), 169–188. doi: 10.1023/A:1020363010465.

Rutkowski, L., Gonzalez, E. J., Joncas, M. & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151. doi: 10.3102/0013189X10363170.

Schleicher, A. (2014). *Equity, excellence and inclusiveness in education: Policy lessons from around the world.* Paris: OECD Publishing.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S. & Nores, M. (2005). *Lifetime effects: The high/scope perry preschool study through age 40.* Ypsilanti, MO: High/Scope Press.

Stuart, E. A. (2007). Estimating causal effects using school-level datasets. *Educational Researcher, 36*(4), 187–198. doi: 10.3102/0013189X07303396.

Stuart, E. A. (2010). Matching methods for causal inference: A review and look forward. *Statistical Science, 25*(1), 1–21. doi: 10.1214/09-STS313.

Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I. & Taggart, B. (2008). *The effective provision of pre-school education (EPPE) project: Final report from the primary phase: Pre-school, school and family influences on children's development during key stage 2 (Age 7–11).* Retrieved from the Institute of Education, University of London website: http://www.ioe.ac.uk/End_of_primary_school_phase_report.pdf.

UNESCO. (2006). *Strong foundations: Early childhood care and education.* Paris: UNESCO.

Vandenbroeck, M., De Visscher, S., Van Nuffel, K. & Ferla, J. (2008). Mothers' search for infant child care: The dynamic relationship between availability and desirability in a continental European welfare state. *Early Childhood Research Quarterly, 23*(2), 245–258. doi: 10.1016/j.ecresq.2007.09.002.

Vogel, C. A., Xue, Y., Moiduddin, E. M., Carlson, B. L. & Kisker, E. E. (2010). *Early Head Start children in grade 5: Long-term follow-up of the Early Head Start Research and Evaluation Study sample* (OPRE report no. 2011-8). Retrieved from the Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services website: http://www.acf.hhs.gov/programs/opre/resource/early-head-start-children-in-grade-5-long-term-followup-of-the-early-head.

Zachrisson, H. D., Janson, H. & Nærde, A. (2013). Predicting early center care utilization in a context of universal access. *Early Childhood Research Quarterly, 28*(1), 74–82. doi: 10.1016/j.ecresq.2012.06.004.