Jan Spilski, Jochen Mayerl, Kirstin Bergström, Maria Klatte

28.02.2015

NORAH-Kinderstudie

Sensitivitätsanalysen Lesemodelle – Fluglärmexposition Schulstandort und Wohnort

Neben den Modellberechnungen wurden zusätzliche Prüfungen der statistischen Validität vorgenommen. Die Prüfungen wurden im Rahmen von Sensitivitätsanalysen durchgeführt.

Um sicherzustellen, dass die im Bericht zur NORAH-Kinderstudie (Klatte, Bergström, Spilski, Mayerl & Meis, 2014) dargelegten Ergebnisse der hierarchischen linearen Modellierung auch bei einer maximalen Gleichverteilung von Merkmalsausprägungen auf Drittvariablen in den verschiedenen Fluglärmpegelgruppen stabil bleiben, wurde ein „Propensity Score Matching“ (PSM) durchgeführt. Anschließend wurden die Mehrebenen-Modelle mit der anhand der Propensity Scores „gematchten“ Stichprobe berechnet. Im Folgenden wird zunächst das Verfahren PSM als Matching für zwei Gruppen erläutert. Im Anschluss wird die Stabilität der Modellösungen für das Leseverständnis mit dem Prädiktor „Fluglärm (Schule)“ ($\mathbf{L}_{pAS,eq,08\text{-}14}$) und mit dem Prädiktor „Fluglärm (Wohnort)“ ($\mathbf{L}_{pAS,eq,06\text{-}18}$) gezeigt. In der Darstellung der Modellösung wird das „volladjustierte Modell“ der NORAH-Stichprobe den Matching-Stichproben gegenüber gestellt, sodass ein Vergleich mit den Ergebnissen des Berichts zur NORAH-Kinderstudie (Klatte, Bergström, Spilski, Mayerl & Meis, 2014) möglich ist.

1. Propensity score matching (PSM)

Oft unterscheiden sich Untersuchungsgruppen nicht nur hinsichtlich der interessierenden unabhängigen Variablen (z.B. Fluglärmexposition), sondern zusätzlich hinsichtlich anderer Merkmale, die einen Einfluss auf die abhängige Variable (Outcome-Variable) ausüben können. Die Gefahr einer solchen Überlagerung von Einflussfaktoren besteht insbesondere dann, wenn aufgrund praktischer oder ethischer Gründe keine „echte“ Randomisierung (Zufallszuweisung der Probanden zu den Untersuchungsgruppen) erfolgen kann. Im Falle solcher Unterschiede sind die Gruppen nicht mehr vergleichbar, und die Ergebnisse bezüglich des Einflusses der unabhängigen Variablen würden verzerrt (Rubin, 2005). Das PSM wird häufig in quasi-experimentellen Studien eingesetzt, um derartige Stichprobenverzerrungen nachträglich statistisch zu kontrollieren. Matching-Ansätze, insbesondere das PSM, sind in der Psychologie und Bildungswissenschaft aber noch wenig verbreitet, obwohl ein exponentieller Anstieg von Veröffentlichungen stattfindet (siehe

dazu Abbildung 1). Die vergleichsweise geringe Anwendung resultiert zum Teil aus der Erhebungsmethodik von psychologischen und bildungswissenschaftlichen Studien. In diesen Studien wird oft bereits a priori eine Parallelisierung (direct matching) der Untersuchungsgruppen hinsichtlich wesentlicher Einflussfaktoren auf die abhängige Variable realisiert, um eine mögliche Konfundierung bereits im Vorfeld auszuschließen. Das a priori „direct matching“ wurde im Rahmen der NORAH-Kinderstudie auch erfolgreich umgesetzt und in Kapitel 2.2.3 „Kriterien der Schulauswahl“ beschrieben (Klatte et al., 2014, S. 45). Bei der nachträglichen statistischen Korrektur ist jedoch das PSM das momentan am häufigsten eingesetzte Matching-Verfahren (Lipowsky et al., 2014).

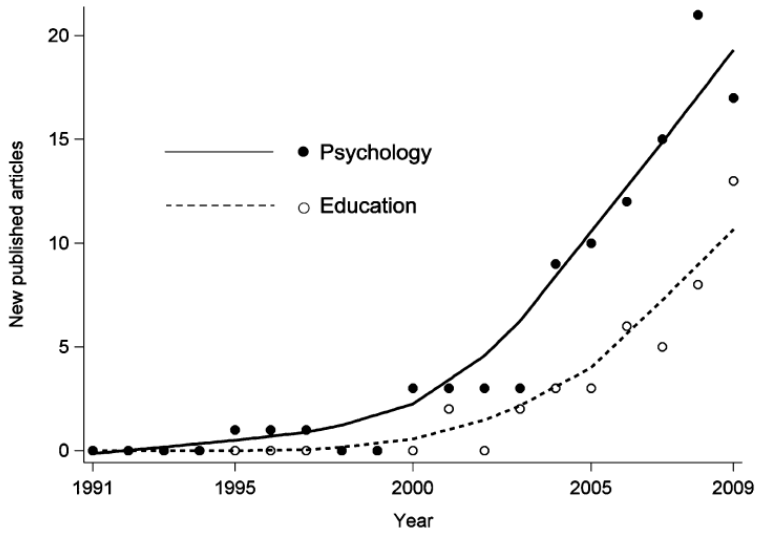

Abbildung 1. Entnommen aus Thoemmes & Kim (2011, S. 91). Entwicklung der Anzahl von Veröffentlichungen mit Propensity Scores in der Psychologie und Bildungsforschung zwischen 1991 und 2009.

1.1. Bestimmung der Propensity Scores

Beim PSM werden im ersten Schritt sogenannte Propensity Scores bestimmt. Diese können durch logistische Regressionsmodelle, probit Regressionsmodelle sowie Diskriminationsanalysen berechnet werden (Rosenbaum & Rubin, 1983; siehe aber auch McCaffrey et al. (2004), die einen Data-mining-Algorithmus vorschlagen).

1.1.1. Propensity Scores allgemein

Propensity Scores sind bedingte Wahrscheinlichkeiten, einer Referenzgruppe (z.B. Referenzgruppe: hohe Fluglärmexposition) anzugehören. Sollten sich die Ausprägungen auf den Drittvariablen (wie z.B. sozioökonomischer Status, Sprachkenntnisse, Geschlecht) zwischen den Gruppen so unterscheiden, dass die Wahrscheinlichkeit für Probanden mit höherer oder geringerer Ausprägung höher ist, in der Referenzgruppe als in der

Kontrollgruppe zu sein, dann würde dies die Validität der Ergebnisse erheblich einschränken. Aufgrund des Ungleichgewichts der Merkmalsverteilungen der Drittvariablen würde eine Konfundierung vorliegen.

Ein PSM ist nur für bereits erhobene Variablen möglich. Eine nachträgliche Korrektur kann somit nur für a priori spezifizierte Drittvariablen durchgeführt werden. Häufig werden jedoch nur wenige Drittvariablen (Kovariaten) in das Matching einbezogen, so dass eine Verzerrung (Konfundierung) auch nach einem Propensity Score Matching gegeben sein kann (Shadish et al., 2006).

Um gleichzeitig Propensity Scores zu bestimmen und ein Matching vornehmen zu können wurde auf das SPSS Erweiterungsskript „Fuzzy“ und eine syntaxbasierte Berechnung zurückgegriffen. Durch die Inspektion der Wald-Statistik ist es möglich, Aussagen darüber zu treffen, ob in der bisherigen Stichprobe signifikante Unterschiede zwischen den Gruppen hinsichtlich der Drittvariablen (z.B. sozioökonomischer Status) bestehen, was ein Indiz für einen Selektionsbias sein könnte. Im „besten“ Fall sollte sich bei zwei Gruppen aufgrund der aufgenommenen Drittvariablen nur in 50% der Klassifikationen eine korrekte Zuordnung zu den Untersuchungsgruppen ergeben, mit einer Verteilung der Klassifikationshäufigkeiten von 25% in jeder der vier Zellen (siehe Tabelle 1). Eine solche Verteilung bedeutet, dass die Wahrscheinlichkeit, aufgrund der Ausprägungen in den Drittvariablen einer der beiden Expositionsgruppen („geringe Fluglärmexposition“ vs. „hohe Fluglärmexposition“) anzugehören, je 50 % beträgt. Die Ausprägung der Drittvariablen ermöglicht somit keine Vorhersage über die Gruppenzugehörigkeit.

Eine weitere Möglichkeit, ein PSM durchzuführen und Propensity Scores zu bestimmen, bietet das Statistikpaket R. Die Berechnung der Propensity Scores (logistische Regressionsmodelle) erfolgt mit dem R-Package MatchIt (Ho et al., 2007).

Tabelle 1. Klassifikationstabelle (exemplarisch)

		Vorhersagewert		
		Fluglärmexposition: niedriger = 0 und höher = 1		Prozentsatz richtig
Beobachtet		0	1	
Fluglärmexposition (niedriger = 0 und höher = 1)	0	***25***	25	25
	1	25	***25***	25
Gesamtprozentsatz				50

Anmerkungen. Die kursiv hervorgehoben Werte stehen für korrekte Zuweisungen. Hier 25 Personen bei *Ngesamt* = 100 entspricht das auch 25% der Gesamtstichprobe. Diese 25% wurden aufgrund ihrer Ausprägungen in den Drittvariablen korrekt der Expositionsgruppe (Fluglärm geringer = 0) zugeordnet. Gleichzeitig wurden aber auch 25% falsch zugeordnet. Damit liegt die Wahrscheinlichkeit bei 50 %, in der einen oder der anderen Expositionsgruppe zu sein.

1.1.2. Propensity Scores - Sensitivitätsanalysen NORAH

Die Bestimmung der Propensity Scores erfolgte im Rahmen der NORAH-Studie durch die Berechnung logistischer Regressionsanalysen. Dafür wurde eine Stratifizierung in zwei und drei Fluglärm-Pegelklassen am Schulstandort bzw. am Wohnort durchgeführt. Diese Pegelklassen wurden als abhängige Variablen und die Kontrollvariablen (Kovariaten) der Mehrebenenmodelle als unabhängige Variablen in die logistischen Modelle aufgenommen. In Tabelle 2 sind die entsprechenden Kovariaten für die Lesemodelle aufgeführt. Die Propensity Scores basieren immer auf Level 1-Prädiktoren. Die Wald-Statistik signalisierte nur in 20% der Fälle und dann jeweils nur bei einer von 10 Kovariaten pro Analyse signifikante Unterschiede zwischen den Gruppen. Eine Systematik war nicht erkennbar, da bei jedem neuen Matching eine andere Kovariate ein signifikanter Prädiktor war. Inspektionen der Klassifikationstabellen zeigten, dass die Wahrscheinlichkeiten von Probanden, aufgrund der aufgenommenen Kovariaten z.B. einer von zwei Expositionsgruppen (niedrige vs. höhere Fluglärmexposition) anzugehören, um 50% streuten (maximal 57%). Das bedeutet, dass die Wahrscheinlichkeit in über 90% der logistischen Modelle bei annähernd 50:50 lag. Daher ist nach diesem Schritt bereits erkennbar, dass der a priori gewählte direct Matching-Ansatz (siehe S.45 NORAH-Endbericht, Klatte et al., 2014) bei der Datenerhebung funktioniert hat: Die Merkmalausprägungen verteilen sich bereits ohne PSM sehr ähnlich auf die Fluglärm-Pegelklassen.

Tabelle 2. Kovariaten für die Lesemodelle

	Lesemodelle L1-Ebene
	Alter
	Geschlecht (0=m, 1=w)
	SWI (Haushalt)
Kovariaten	Migrationshintergrund
	Deutschrating
	Anzahl Kinderbücher
	Nichtsprachliche Fähigkeiten
	Auditives Gedächtnis
	Bildertest
	Phonologische Bewusstheit

1.2. Matching

1.2.1. Matching allgemein

Es gibt unterschiedliche Matching-Strategien um nachträglich einen möglichen Selektionsbias statistisch zu kontrollieren. Exemplarisch werden einige aufgeführt. Das 1:1 Matching ist eine Form des statistischen Matching, bei dem ein Proband einem exakten oder annähernd gleichen Zwilling zugeordnet wird. Beim „one-to-many Matching“ wird ein Proband mit z.B. zwei anderen Probanden gematcht, so dass hier Drillinge vorliegen würden. Es sind aber auch Vierlinge und mehr möglich (Ming & Rosenbaum, 2000). Weitere Matching-Formen sind das „Full-Matching“ (Hansen, 2004), das „nearest neighbor-Matching“ (Rosenbaum & Rubin, 1985) sowie das „Kernel-Matching“ (Heckmann et al., 1997). Neben den eben genannten nachträglichen statistischen Matching-Verfahren ist aber eine Parallelisierung der Untersuchungsgruppen durch ein a priori direct-Matching möglich und sinnvoll, wie es in der NORAH-Kinderstudie auch umgesetzt wurde.

1.2.2. Matching – Sensitivitätsanalysen NORAH

Im Rahmen der Sensitivitätsanalyse der NORAH-Kinderstudie wurde das „nearest neighbor-Matching“ verwendet. Das Matching basierte hierbei immer auf den Propensity Scores, die durch logistische Regressionsmodelle mit den Level 1-Kovariaten geschätzt wurden. Ein Matching für Level 2-Kovariaten konvergierte zu keiner Lösung, so dass der iterative Prozess abgebrochen wurde. Aus diesem Grund wurde kein Matching für die Level 2-Kovariaten durchgeführt.

Beim „nearest neighbor-Matching“ wurde für jeden Probanden der Referenzgruppe (hohe Fluglärmexposition am Schulstandort bzw. Wohnort) ein Proband aus einer anderen Gruppe (geringere Fluglärmexposition am Schulstandort bzw. Wohnort) mit vergleichbarem Propensity Score zugeordnet. Probanden, denen kein „Zwilling“ mit vergleichbarem Propensity Score zugeordnet werden konnte, gingen in die anschließenden Berechnungen nicht mehr ein. Rosenbaum & Rubin (1985) empfehlen als Kriterium der Distanzreduktion zwischen zwei Probanden eine Standardabweichung (*SD*) von 0,25. Neuere Studien (Austin, 2009; Wang, Li, Jiang et al., 2013) empfehlen jedoch ein Kriterium von 0,20. Bei den vorliegenden Analysen wurde den neueren Studien gefolgt und das Kriterium *SD* = 0,20 zugrunde gelegt.

Das Matching erfolgte zum einen auf die durch *Mediansplit* geschaffenen Gruppen niedrigere vs. höhere Fluglärmexposition am Schulstandort bzw. am Wohnort, und zum anderen auf die im NORAH-Bericht (vgl. Klatte et al., 2014, S. 96f.) spezifizierten

Fluglärm-Expositiongruppen „geringe Exposition“ (< 47 dB) und „hohe Exposition“ (>= 55 dB) (Extremgruppen-Matching).

Die Stichprobe reduzierte sich durch das Matching von ursprünglich *N* = 1090 auf *n* = 1006 beim Median-Matching (Matching der Probanden mit Fluglärmexposition an Schule bzw. Wohnort unter vs. über dem Median) und auf *n* = 708 beim Extremgruppen-Matching (Matching der gering mit der hoch exponierten Gruppe). Die Propensity Scores zeigten in allen Substichproben große Überlappungen, was als ein Indiz für ein adäquates apriori-Matching angesehen werden kann (siehe exemplarisch die Verteilung der PSM in Abbildung 2). Extremwerte mit Propensity Scores in der Nähe von 0 oder 1 traten ebenfalls nicht auf. Daher kann festgehalten werden, dass eine Verzerrung der Schätzungen der Mehrebenenanalysen aufgrund von Selektionseffekten als sehr gering angesehen werden kann.

In Tabelle 3 und 4 sind die mehrebenenanalytischen Auswertungen der gematchten und der nicht gematchten (ursprünglichen) Stichproben gegenüber gestellt. Es wird deutlich, dass sich die Modelle auch bei einer reduzierten und gematchten Stichprobe weder in der Richtung des Zusammenhangs (Steigungskoeffizient) noch in den Intercepts unterscheiden. Des Weiteren kann festgehalten werden, dass Aussagen über die Signifikanz auch bei den reduzierten Matching-Stichproben wie in der ursprünglichen Stichprobe beibehalten werden können. Die Übereinstimmung der Ergebnisse beider Verfahren ist konsistent mit den Ergebnissen von Metaanalysen, in denen Propensity Score-Methoden mit traditionellen regressionsanalytischen Verfahren verglichen wurden (Shah et al., 2005; Stürmer et al., 2006).

Auch durch die nachträgliche statistische Kontrolle möglicher Stichprobenverzerrungen ist es möglich, die bisherigen Aussagen der NORAH-Kinderstudie bezüglich des Zusammenhangs zwischen Fluglärmexposition und verringerten Leseleistungen aufrecht zu erhalten. Die im Bericht dargelegten statistischen Modelle können somit als robust betrachtet werden.

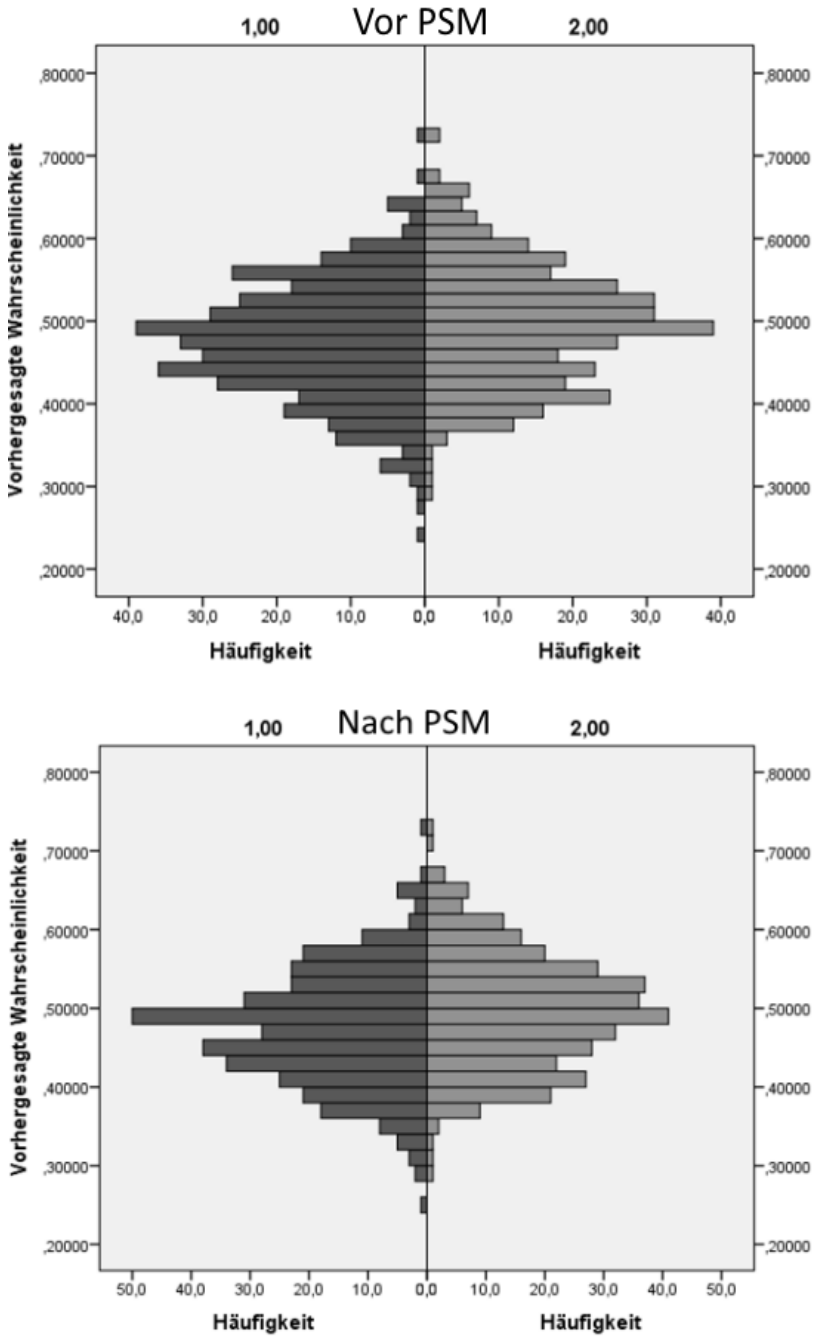

Abbildung 2. Verteilung der Propensity Scores vor (oberer Teil der Abbildung) und nach dem propensity score matching (PSM), für die Gruppen „geringe Exposition Schulstandort“ im linken Teil der Abbildung (1,00) und „hohe Exposition Schulstandort“ im rechten Teil der Abbildung (2,00).

Tabelle 3. *Modellparameter der Mehrebenenanalysen für Outcome-Variable „Leseverständnis, Prädiktor* ***„Fluglärm (Schule)“*** *($L_{pAS,eq,08-14}$) und Kontrollvariablen, Gesamtgruppe, Median- und Extremgruppen Propensity Score Matching (PSM)*

	Endmodell (volladjustiert)		Median (PSM) Match		Extremgruppen (PSM) Match	
	b (SE)	*p*	*b (SE)*	*p*	*b (SE)*	*p*
Wortverständnis						
N=1090/1006/708						
ICC=.091/.057/.087						
Intercept	46,45 (0,640)		46,27 (0,646)		46,06 (0,792)	
Fluglärm (Schule) – Level 2	-0,105 (0,064)	0,049	-0,128 (0,062)	0,020	-0,130 (0,072)	0,035
Satzverständnis						
N=1090/1006/708						
ICC=.090/.036/.045						
Intercept	45,03 (0,543)		44,80 (0,541)		44,80 (0,644)	
Fluglärm (Schule) – Level 2	-0,064 (0,056)	0,125	-0,091 (0,056)	0,053	-0,065 (0,058)	0,130
Textverständnis						
N=1090/1006/708						
ICC=.062/.018/.028						
Intercept	46,33 (0,570)		46,14 (0,587)		45,79 (0,668)	
Fluglärm (Schule) – Level 2	-0,118 (0,045)	0,005	-0,133 (0,044)	0,002	-0,124 (0,051)	0,008
Gesamttest						
N=1090/1006/708						
ICC=.081/.035/.052						
Intercept	45,94 (0,534)		45,74 (0,539)		45,63 (0,644)	
Fluglärm (Schule) – Level 2	-0,097 (0,050)	0,027	-0,119 (0,049)	0,007	-0,107 (0,056)	0,029

SE = Standardfehler (standard error). Signifikanz (*p-Werte)* sind für die einseitige Hypothesenprüfung angegeben.

Tabelle 4. *Modellparameter der Mehrebenenanalysen für Outcome-Variablen „Leseverständnis", Prädiktor* ***„Fluglärm (Wohnort)"*** *($L_{pAS,eq,06-18}$) und Kontrollvariablen, Gesamtgruppe, Median- und Extremgruppen Propensity Score Matching (PSM)*

	Endmodell (volladjustiert)		Median (PSM) Match		Extremgruppen (PSM) Match	
	b (SE)	*P*	*b (SE)*	*p*	*b (SE)*	*p*
Wortverständnis						
N=1089/933/552						
ICC=.091/.055/.081						
Intercept	46,94 (1,252)		46,52 (1,062)		46,67 (1,051)	
Fluglärm (Wohnort) – Level 1	-0,078 (0,053)	0,069	-0,074 (0,056)	0,094	-0,131 (0,075)	0,041
Satzverständnis						
N=1089/936/552						
ICC=.090/.041/.026						
Intercept	45,44 (1,173)		45,29 (0,990)		45,09 (0,953)	
Fluglärm (Wohnort) – Level 1	-0,057 (0,054)	0,144	-0,061 (0,054)	0,131	-0,099 (0,067)	0,070
Textverständnis						
N=1089/936/552						
ICC=.091/.055/.014						
Intercept	47,12 (1,107)		47,18 (0,934)		47,57 (0,953)	
Fluglärm (Wohnort) – Level 1	-0,096 (0,042)	0,011	-0,103 (0,043)	0,008	-0,102 (0,057)	0,039
Gesamttest						
N=1089/936/552						
ICC=.081/.037/.039						
Intercept	46,54 (1,048)		46,34 (0,904)		47,11 (1,803)	
Fluglärm (Wohnort) – Level 1	-0,080 (0,044)	0,036	-0,083 (0,046)	0,034	-0,111 (0,063)	0,041

SE = Standardfehler (standard error). Signifikanz (*p-Werte)* sind für die einseitige Hypothesenprüfung angegeben.

Quellenverzeichnis

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

Bootzin, R. R., & McKnight, P. E. (Eds.). (2006). *Strengthening research methodology: Psychological measurement and evaluation*. Washington: American Psychological Association.

Bryer, J. (2013). TriMatch R-package. Retrieved from http://jason.bryer.org/TriMatch

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American Statistical Association, 99*(467), 609–618. doi:10.1198/016214504000000647

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies, 65*(2), 261–294.

Klatte, M., Bergström, K., Spilski, J., Mayerl, J. & Meis, M. (2014). NORAH Noise-related annoyance, cognition, and health: Wirkungen chronischer Fluglärmbelastung auf kognitive Leistungen und Lebensqualität beiGrundschulkindern. URL http://www.laermstudie.de/ergebnisse/ergebnisse-kinderstudie/ueberblick/

Lipowsky, F., Stubbe, T. C., Faust, G., Künsting, J., Hadeler, S., & Bos, W. (2014). Was leisten Schülerinnen und Schüler der privaten BIP-Kreativitätsgrundschulen im nationalen Vergleich. *Journal für Bildungsforschung Online, 6*(2), 89–112.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods, 9*(4), 403.

Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56*(1), 118–124.

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.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician, 39*(1), 33–38. doi:10.1080/00031305.1985.10479383

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association, 100*(469), 322–331. doi:10.1198/016214504000001880

Shah, B.; Laupacis, A.; Hux, J.; Austin, P. (2005). Propensity score modeling gave similar results to traditional regression modeling in observational studies: A systematic review. Journal of Clinical Epidemiology, 58, 550-559.

Shadish, W. R., Luellen, J. K., & Clark, M. H. (2006). Propensity Scores and Quasi-Experiments: A Testimony to the Practical Side of Lee Sechrest. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143–157). Washington: American Psychological Association.

Stürmer, T.; Joshi, M.; Glynn, R.; Avorn, J.; Rothman, K.; Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but no substantially different estimates compared with conventional multivariate methods. Journal of Clinical Epidemiology, 59, 437-447.

Thoemmes, F. J., & Kim, E. S. (2011). A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research, 46*(1), 90–118. doi:10.1080/00273171.2011.540475

Wang, Y., Cai, H., Li, C., Jiang, Z., Wang, L., Song, J., . . . Hills, R. K. (2013). Optimal Caliper Width for Propensity Score Matching of Three Treatment Groups: A Monte Carlo Study. PLoS ONE, 8(12), e81045. doi:10.1371/journal.pone.0081045