Adjusting expectations or maintaining first impressions? The stability of teachers' expectations of students' mathematics achievement

Timmermans, Anneke; Rubie-Davies, Christine M.; Wang, Shengnan

Published in:
Learning and Instruction

DOI:
10.1016/j.learninstruc.2021.101483

Publication date:
2021

Learning and Instruction 75 (2021) 101483

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Adjusting expectations or maintaining first impressions? The stability of teachers' expectations of students' mathematics achievement

Anneke C. Timmermans [a,*], Christine M. Rubie-Davies [b], Shengnan Wang [b]

[a] GION Education/Research, University of Groningen, Grote Rozenstraat 3, 9712 TG, Groningen, the Netherlands
[b] Faculty of Education and Social Work, The University of Auckland, 74 Epsom Ave, Epsom, Auckland, 1023, New Zealand

ARTICLE INFO

Keywords:
Teacher expectations
Stability
Student achievement
Mathematics
Random-intercept crossed-lagged panel models

ABSTRACT

The issue of teacher expectation stability is crucial in understanding the self-fulfilling prophecies generated by teacher expectations. However, currently there is a lack of empirical evidence related to teacher expectation stability. The aim of the current study was to assess the temporal stability of teacher expectations of their students' mathematics achievement within the timeframe of one school year. Random-Intercept Cross-Lagged Panel Models were employed based on a sample of 2536 students taught by 89 teachers in New Zealand elementary and middle public schools. Strong rank order stability was found in teacher expectations at the between-student level. Expectation instability was present at the within-student level. Paths from student mathematics achievement to teacher expectations were stronger than the paths in the opposite direction, indicating that teachers adapted their expectations for students to fall in line with student performance and continued to do so throughout the year.

1. Introduction

In 2018, we celebrated fifty years of research on teacher expectations. Teacher expectations are "primarily cognitive phenomena, inferential judgments that teachers make about probable future achievement and behavior based upon the student's past record and his present achievement and behavior" (Brophy & Good, 1974, p. 129). Studying teacher expectations began with the publication of the influential book "*Pygmalion in the classroom*" by Rosenthal and Jacobson (1968). This ground-breaking study marked the beginning of a rich and flourishing research field investigating teachers' expectations of their students' learning and achievement. In first part of the introduction section, we will introduce the teacher expectation framework by providing a brief overview of the major directions of the existing research, after which we will introduce and discuss the research gap that will be addressed in the current study.

Previous research has shown that teachers are relatively accurate in their expectations (Jussim & Harber, 2005), but nevertheless favor some students over others in their expectations (e.g., De Boer, Bosker, & van der Werf, 2010; Glock & Krolak-Schwerdt, 2013; Tenenbaum & Ruck, 2007; Timmermans, Kuyper, & Van der Werf, 2015). Teachers' expectations affect subsequent teaching behaviors; for example, through asking richer questions, and providing learning-focused feedback to students for whom the teachers have high expectations (e.g., Brophy & Good, 1970; Good & Lavigne, 2018; Rubie-Davies, 2007; Weinstein, 2002). Teacher expectations work as a self-fulfilling prophecy (Merton, 1948) on subsequent student outcomes such as academic performance, intelligence, self-efficacy, and motivation (e.g., Agirdag, Van Avermaet, & Van Houtte, 2013; McKown & Weinstein, 2008; Tyler & Boelter, 2008; Zhu & Urhahne, 2015). Recent evidence has indicated that the previous findings are not universal, however, as some students, including low achievers, students from low-income families, and those from ethnic minority groups (e.g., Hinnant, O'Brien, & Ghazarian, 2009; McKown & Weinstein, 2002; 2008), seem to be more susceptible to self-fulfilling prophecy effects of teacher expectations than other students. Moreover, some teachers place more credence in student differences than others (e.g., Rubie-Davies, 2007; Timmermans, de Boer, & van der Werf, 2016), and a small group of teachers seems to generate stronger self-fulfilling prophecy effects on students' academic achievement than the majority of teachers (e.g., Weinstein, 2002; Rubie-Davies, 2015; Timmermans & Rubie-Davies, 2018).

To date, however, very little research has been available indicating whether the expectations held by teachers for individual students are stable over time or whether first impressions tend to get adjusted when

* Corresponding author. University of Groningen, GION Education/Research, Grote Rozenstraat 3, 9712 TG, Groningen, the Netherlands.
E-mail address: A.C.Timmermans@rug.nl (A.C. Timmermans).

https://doi.org/10.1016/j.learninstruc.2021.101483
Received 30 March 2020; Received in revised form 12 March 2021; Accepted 20 March 2021
Available online 21 April 2021

new information becomes available (Wang, Rubie-Davies, & Meissel, 2020). Teacher expectation research has relied almost exclusively on studies with single teacher expectation measurements that are taken early in the academic year (Rubie-Davies, Watson, Flint, Garrett, & McDonald, 2018). The stability of teacher expectations refers to "whether or not teachers maintain their existing expectations in the face of contradictory information or whether they adjust their expectations as students progress through the year" (Rubie-Davies et al., 2018, p. 223). Although this gap in empirical evidence has been recognized since the early 1980s (e.g., Brophy, 1983; Martinek, 1980), and some studies have alluded to the idea of stability of teacher expectation effects (e.g., De Boer et al., 2010; Hinnant et al., 2009; Rubie-Davies et al., 2014), the field has made only small steps forward in empirically investigating teacher expectation stability within the course of a school year.

The issue of teacher expectation stability is, however, crucial in understanding the self-fulfilling prophecies they generate. For self-fulfilling prophecies to occur, the teachers' initial expectations must be inaccurate (Merton, 1948), and the teacher must sustain these inaccurate beliefs and act accordingly (Brophy, 1983). Therefore, some degree of instability in teacher expectations may indicate that teachers adjust or recalibrate their original expectations and that their expectations may be subject to changes in student classroom behaviors and/or academic performance of students. As teachers gradually obtain more information about their students, and if they recalibrate their expectations to fall in line with the obtained information on students, it is likely that the gap between teacher expectations and student achievement becomes smaller (i.e., teachers become more accurate in expectations as time passes). Moreover, if teacher expectations become more accurate over time, the potential for teacher expectations to function as self-fulfilling prophecies would be reduced (Jussim & Harber, 2005). On the other hand, if teacher expectations are relatively stable across time and do not respond to updated evidence and information about students, it would be more likely for self-fulling prophecy effects to occur. In that scenario there is the possibility that the expectation effects would accumulate rather than dissipate over a sustained period (Wang et al., 2020).

To sum up, evidence on whether teacher expectations remain stable over a relatively long time period could provide useful insights into understanding several essential yet controversial issues in the teacher expectation field, which include the accuracy of teacher expectations (Brophy, 1983), the accumulation or dissipation of expectation effects (Jussim & Harber, 2005), and the magnitude of expectation effects on influencing student achievement. A deeper understanding of these issues could have important implications for instructional practices and student achievement. The aim of the current study was to assess the temporal stability of teacher expectations within the timeframe of one school year.

1.1. Teacher expectation stability views

There are two competing views concerning the potential (in)stability of expectations (Levine & Doyle, 2002; Wang et al., 2020). Some researchers have contended that teachers adjust their expectations for students to fall in line with achievement (Brophy, 1983), indicating that expectations are unstable to at least some degree. New information indicating that the original expectations were either too high or too low then leads to a correction. This view is consistent with a social constructivist perspective, which emphasizes the classroom environment as an evolving entity that is continually reshaped by both the teacher and the students (Darley & Fazio, 1980; Rosenholtz & Simpson, 1984). However, if teachers do adjust their original expectations, there may still be self-fulfilling prophecies of teacher expectations on subsequent student performance albeit smaller than if the teacher had sustained an initial inaccurate expectation because expectations and performance quickly converge. This process of adaptive expectations would lead to a reciprocal relationship between teacher expectations and student performance in which the teachers' expectations shaped student performance by means of self-fulfilling prophecies and student performance shaped the teachers' expectations through adjustments or corrections (Raudenbush, 1984).

Other researchers have noted that people are likely to adhere to their initial expectations (Babad, 2009; Kuklinski & Weinstein, 2000; Rubie-Davies et al., 2018), therefore suggesting strong stability in expectations. This would imply that teachers tend to stick to their initial beliefs or first impressions about student aptitude, whether they are correct or not. Adaptations of or corrections to expectations are less likely to happen, implying that expectations and performance converge at a slower rate. The process of stable expectations could lead to a unidirectional relationship between expectations and performance in which expectations shape student performance by means of self-fulfilling prophecies.

There may be various conditions under which expectations are likely to be maintained despite disconfirming evidence, as cognitive biases can exert a sustaining influence on a teachers' initial expectations (Jussim, 1986). Ambiguous information about a student's performance can be perceived by teachers in expectancy-confirming ways, discrepancies between actual behavior and expectations can be discounted easily, and expectancy-consistent information may be remembered more easily. For example, expectancy-consistent performance is often attributed to the personal characteristics of students (e.g., intelligence, gender), whereas expectancy-inconsistent performances are attributed to factors outside the students, such as situational factors (e.g., luck; Deaux & Emswiller, 1974; Regan, Straus, & Fazio, 1974). Moreover, the frequency of cases in which performance is consistent with expectations tends to get overestimated (Chapman, 1967) and confirming cases are more easily recalled (Crocker, 1981).

1.2. Empirical evidence regarding teacher expectation stability

In this section, we elaborate on the results of empirical research in which the same sample of teachers provided information on their expectations of individual students on multiple occasions within the course of a school year, thus excluding studies of how teacher expectations are transferred from one teacher to the next (e.g., Hinnant et al., 2009; Rubie-Davies et al., 2014) or whether the effects of teacher expectations measured at one point in time are stable or dissipate over the longer term (e.g., Alvidrez & Weinstein, 1999; De Boer et al., 2010; Sorhagen, 2013).

To our knowledge, only six studies could be identified in which teacher expectation stability has been empirically studied. An overview of these studies, their main characteristics, and outcomes is presented in Table 1. In order to interpret the stability found in previous research we used the guidelines presented by Wang et al. (2020): strong correlational stability $r > 0.70$, moderate correlational stability $0.7 > r \geq 0.5$, some stability $0.5 > r \geq 0.3$, little or low stability $r < 0.3$.

Three studies investigated teacher expectation stability within a single school year in primary education using two (Kuklinski & Weinstein, 2000; Martinek, 1980) or three measurements (Good, Cooper, & Blakey, 1980), but with varying time lags ranging from 8 weeks (Martinek, 1980) to approximately 6 months (Kuklinski & Weinstein, 2000) between the measurements of expectations. These studies, investigating various subject domains including physical education skills (Martinek, 1980) and reading (Kuklinski & Weinstein, 2000), all indicated moderate to strong relative stability of teacher expectations over time (see Table 1). For example, the correlations between teacher expectations regarding the students' general academic potential and success at completing verbal academic tasks over time lags of three to four months were all above 0.75 (Good et al., 1980). This implies that the students for whom teachers had high expectations on one occasion were likely the same students for whom the teachers had high expectations four months later. Comparable indications of stability were found by Clifton (1981) in secondary education. When groups of teachers had to form

Table 1
Overview of previous teacher expectation stability papers.

Publication	#Measurements	Type of measurement	Time lag	Sector	Domain	Control variables	#Teacher	#Student	Reported stability (*r* or otherwise specified)
Martinek (1980)	2	Scale (not further specified)	8 weeks	Primary education	Overall performance in physical skill, social relations, cooperative behaviour, ability to reason	No controls	6	179	.96 - .84 for overall skill; .96 - .83 for social relations; .93 - .78 for cooperation; .92 - .68 for reasoning
Good et al (1980)	3	1 item per outcome, ranking students	3-4 months	Primary education	General academic potential, success at verbal academic tasks	No controls	16	192	All correlations exceeded .75 (no further specification)
Clifton (1981)	4	1 item per outcome, 5-point rating scale (much below average – much above average)	1 year	Secondary education	Reliability, cooperation with teachers and students, industry in schoolwork, successfully completing high school	No controls	?	6158	.71 for reliability; .69 for cooperation; .74 for industry; .85 for completing high school
Kuklinski and Weinstein (2000) Sample 1	2	1 item, ranking students 1 item, 5 point rating scale (poor – outstanding)	6 months	Primary education	Reading	No controls	48	464	Median tau .69
Kuklinski and Weinstein (2000) Sample 2	2	1 item, ranking students	6 months	Primary education	Reading	No controls	12	138	Median tau .65
Rubie-Davies et al. (2018) Year 1	3	1 item per outcome, 7 point rating scale (very much below average – very much above average)	4 months	Primary and intermediate schools	Reading and mathematics	Reading and mathematics achievement	94	?	No significant changes in means (associated with very small effect sizes)
Wang et al. (2020)	3	1 item (expectations school-year final examination, 13 levels)	3-5 months	Secondary education	Chinese, mathematics and English	Chinese, mathematics and English achievement	48	1199	.15 - .30 Chinese (mean Rho) .23 - .34 mathematics (mean Rho) .22 -.40 English (mean Rho)

expectations for individual students in secondary education, these expectations proved relatively stable even with time lags of one year.

It should be noted that the aforementioned teacher expectation stability studies did not control for students' actual achievement levels at the time of measurement of the expectations. Therefore, high correlations in the teachers' expectations between time points were to be expected as previous research has shown strong stability in students' academic performance (e.g., Marks, 2016; Timmermans & van der Werf, 2018) and a relatively high correspondence of teacher expectations with students' actual academic performance (e.g., Jussim & Harber, 2005; Timmermans, van der Werf, & Rubie-Davies, 2019). Therefore, these correlations between teacher expectations for individual students over multiple time points may be a stronger reflection of stability in student academic performance rather than indications of whether teachers are able to adjust their expectations. This also becomes clear from the study of Wang et al. (2020) in which teacher expectation stability was investigated for Chinese high school students. After controlling for current performance levels at each of the three measurement occasions, correlations between teacher expectations with a three-to five-month gap were much lower, generally pointing to little or some stability. In only 18–22 percent of the classes, was teacher expectations stability moderate or strong.

Kuklinski and Weinstein (2000) also employed another approach to investigating teacher expectation stability. Besides calculating correlations between expectations, they investigated whether the teachers' expectations as measured using ratings (i.e., teachers' ratings of expected students' performance levels in the domain of reading) were similar in the fall and spring. Over half the students (52%) were rated similarly by their teachers in the fall and spring of the school year. If a teacher adjusted their expectations for a student, this was usually by one level (potential range 0–4). Larger adjustments only occurred for about 6% of the students.

Findings concerning stability were recently extended in a study by Rubie-Davies et al. (2018), albeit a study of teacher expectation stability at the teacher level rather than at the level of individual students. In their study, teacher expectations were corrected for students' current achievement levels, and stability was not only studied based on three measurements of expectations within a single school year, but teachers were also surveyed for subsequent cohorts of students twice each year. In this study, if teachers had high expectations for all the students in their class relative to the other teachers in the same sample, they continued to have high expectations for all students in their class over one year and across different cohorts of students. The results showed that class-level teacher expectations were not only stable within the school year, but high expectation teachers continued to have high expectations over longer periods of time. Even when they taught new student groups, their expectations for the new students continued to be relatively high. Similarly, low expectation teachers remained low expectation teachers. Although important, the finding that high expectation teachers remained high expectation teachers, across one year and for subsequent cohorts does not provide a satisfactory answer to the question of whether expectations of individual students are adapted when new performance information becomes available for teachers.

In summary, previous research on teacher expectation stability at the student level has indicated moderate to strong stability when current achievement levels were not taken into account (e.g., Clifton, 1981; Kuklinski & Weinstein, 2000; Martinek, 1980), but little to some stability when current performance levels were taken into account (Wang et al., 2020). At the teacher level, expectations remained relatively stable across one year and for subsequent cohorts (Rubie-Davies et al., 2018).

1.3. Methodology for investigating teacher expectation stability

In previous studies, teacher expectation stability was usually considered in terms of consistency in teachers' ratings of students' performance levels or rank ordering of students in relation to their future achievement over multiple time points (Kuklinski & Weinstein, 2000), and this was usually assessed using measures of association. In these studies, stability was therefore considered a between-student phenomenon (i.e., does the ranking of students in teacher expectations change over time?). This does not align with the idea that instability in expectation refers to teachers adjusting their expectations for individual students (i.e., does the level of teacher expectations for a particular student change over time?). Adjusting expectations to fall in line with new information about a student would imply stability as, at least partially, a within-student phenomenon. Therefore, it seems important to distinguish the consistency of a teacher's expectations for an individual student over time (that is, absolute stability or within-person stability) from the consistency of individual differences between students in their teachers' expectations over time (that is, relative stability or between-person stability; Gustavsson, Weinryb, Goransson, Pedersen, & Asberg, 1997) and to apply methods that allow for an investigation of stability at the within-student level.

Establishing the temporal stability of teacher behavior (including expectations), and, in particular, methodologically separating absolute from relative stability, has long been considered a problematic endeavor (Doyle, 1977; Rogosa, Floden, & Willett, 1984; Shavelson & Dempsey-Atwood, 1976). To achieve this, we needed to consider the longitudinal associations between two variables, that is, the expectation that a teacher had for an individual student and that student's academic achievement. For simultaneous modelling of the associations between two (i.e., expectations and achievement) or more variables measured at multiple occasions, the cross-lagged panel model (CLPM; Rogosa, 1980) is one of the most familiar models. The autoregressive paths (i.e., the coefficient from one to the next measurement of the same variable) from such models are often used as an indication of stability. However, in the CLPM, these autoregressive coefficients reflect the rank order stability of persons (i.e., changes in the ranking of students in teacher expectations after controlling for prior achievement). Additionally, these models allow for an estimation of cross-lagged paths, that is the extent to which one variable is predictive of the other variable at a later time point. For example, the models can show whether student achievement at time point t is predictive for teacher expectations at time point $t+1$ (the adjustment path) over and beyond the predictive value of teacher expectations at time point t. However, the causal paths from the CLPM tend to be over-estimated as they fail to discriminate between- and within-person variation (Burns, Crisp, & Burns, 2019).

Recent advances in panel modelling, however, allow for a better distinction between the two types of stability. The distinction can be made by applying random intercept cross-lagged panel models (RI-CLPM; Hamaker, Kuiper, & Grasman, 2015), which is a multilevel adaptation to the familiar cross-lagged panel model. In the RI-CLPM, two levels are distinguished that are measurements within students at level 1 and between students at level 2. The random intercept cross-lagged panel model separates the within-person process (absolute stability) from the stable between-person differences (relative stability). An additional advantage of this within-person modeling approach is that all stable (unmeasured) confounding variables are controlled for (Becht et al., 2017).

The RI-CLPM models have recently found their way into applications in the educational and pedagogical context, for example, by investigating reciprocal associations between adolescents' self-concept clarity and their relationship quality with parents and friends (Becht et al., 2017), classroom perceptions and motivation (Ruzek & Schenke, 2019), and self-concept, self-efficacy, and achievement (Burns et al., 2019). For instance, Ruzek and Schenke (2019) called for research aimed to compare students to themselves at an earlier time point to determine whether their own unique and changing classroom perceptions were related to their changing motivation, learning, or behavior (within--person), instead of comparing students to one another to determine whether students' differential perceptions of the classroom were related to differences in their motivation, learning, or behavior (between--person). These papers all started by questioning the results found in previous research investigating between-person associations. Moreover, similar to the current study, Burns et al. (2019) and Ruzek and Schenke (2019) compared reciprocal longitudinal associations with both the CLPM and RI-CLPM, with superior model fit for the RI-CLPM in all cases, thus studying the longitudinal reciprocal associations at the within-student level.

1.4. The current study

The goal of the current study was to investigate the stability of teacher expectations for students in New Zealand elementary and middle schools within the time frame of a school year and for the mathematics domain. We aimed to study teacher expectations with a particular focus on within-student stability. Because of the contradicting views that teachers may or may not adjust their expectations of individual students in light of new achievement information and the very small number of empirical studies that are usually based on between-person methods, we did not specify clear hypotheses in one particular direction.

2. Method

2.1. Context

The New Zealand compulsory education sector is comprised of elementary and high school components. Students attend elementary school from Year 1 to Year 8 (aged 5–12 years), with middle schools catering for Years 7 and 8, and, thereafter, students move to the high school system which caters for Years 9–13. Generally, in New Zealand students attend their local elementary and middle school. Very few travel to another school. Schools are ranked by the government on a 1–10 scale, largely based on socioeconomic information for particular geographic areas but the information also includes data related to the percentage of minority group families in the area. On the basis of this ranking, with '1' being a school in a very poor area and '10' being a school in an affluent area, schools are then inversely funded. Schools in poorer areas receive significantly more funding than schools in middle class areas. All schools in New Zealand are self-governing, meaning that a board comprised of the principal, a staff member, and several community members plays a governance role in the running of the school.

2.2. Sample

The participants in this study were 2536 students taught by 89 teachers in nine elementary-level (Years 4–6; aged 8–10 years) and three middle school-level public schools (Years 7 and 8; aged 11–12 years) in low, medium, and high socioeconomic areas. For the sampling, schools within one geographical area of a large city were purposively divided into schools in high, middle, and low socioeconomic areas. Schools were then randomly selected within those socioeconomic categories and invited to participate in the study, thus providing a representative sample of students from within each socioeconomic group. The school-level response rate was 37.5% (12 of 32 schools). Teachers from those twelve schools then volunteered to be part of the study.

Of the 89 teachers, 28.1% were male. Teachers had a variety of experience in terms of teaching years, ranging from 1 to 5 years (32.6%) to over 25 years (18.0%). Concerning the 2536 students, 51% were boys. The average age of the students was 9.63 years, ranging from 6 years (27 students) to 13 years (11 students). In relation to ethnicity, 49.1% were New Zealand European, 17.1% Māori (the indigenous group), 15.9% Pasifika (originating from one of the Pacific Islands), 14.4% Asian, and 3.6% had another background.

2.3. Instruments

2.3.1. Teacher expectations

Expectations regarding mathematics were measured during the first month of the school year (February). By that time, teachers had access to information about students from their previous teachers, including achievement levels, but they had not yet conducted their own standardized testing of the student group. The teacher expectation survey was also administered in the middle (June) and towards the end (October) of the school year. An overview of the teacher expectation measures relative to the New Zealand school year is presented in Fig. 1. On a 1–7 Likert scale, teachers used a one-item survey to predict the levels they believed each student in their class would reach by the end of the year, ranging from 1 = very much below average to 7 = very much above average. At the end of the year, teachers predicted the levels they expected each of their students to reach by the end of the following year when they would be with a new teacher. All teachers used the entire 1–7 range in indicating their expectations for their individual students. This particular instrument has been used in previous studies as a measure of teacher expectations (e.g., Rubie-Davies et al., 2018) and one-item measures are still very common within teacher expectation research (e.g., Bohlmann & Weinstein, 2013; Chalabaev, Jussim, Sarrazin, & Trouilloud, 2009; Kelly & Carbonaro, 2012; Kuklinski & Weinstein, 2001).

2.3.2. Student achievement

Student achievement in mathematics was measured on three measurement occasions using e-asTTle (Electronic Assessment Tools for Teaching and Learning), a standardized measure commonly used in New Zealand schools. The curriculum strands that were tested were number knowledge, number sense, and algebra. For the purposes of the current study, all tests were 40 min long, and teachers selected the most appropriate level for the various students in their class to complete. That is, teachers decided which tests were most appropriate for the individual students in their class. Item response theory (IRT) scoring procedures (Embretson & Reise, 2000) were used to calibrate items; therefore, students should gain a similar score in their test, no matter which test their teacher has selected. Further, each curriculum level in New Zealand covers two years of schooling, so the potential curriculum range of any test is quite broad. Hence, the standardized scores from tests administered across three measurements could be compared. Because of the pre-calibration of the IRT score values, student scores can be tracked over time and compared across classes, year levels, and schools. The national means for student grade and time of year (means are available for each of four quarters) were subtracted from student scores in order to provide scores that could be meaningfully compared across grade levels. The tests are considered highly reliable ($\alpha = 0.96$, $SE = 15$ points). Descriptive statistics on the teachers' expectations and student achievement are presented in Table 2.

2.4. Procedures

Having gained ethical approval for the study from the institutional committee of the second author, teachers completed their estimates of student achievement in February, June, and October as outlined above. In parallel, students completed their e-asTTle tests. Tests were created by the second author and then couriered to teachers. Once their students had completed the tests, they were couriered back to the researchers who marked them.

2.5. Analytic strategy

First, in order to maintain the possibility of comparing our outcome to previous studies we applied similar strategies to assess teacher expectation stability in a descriptive way. Similarly to Clifton (1981), Good et al. (1980), Kuklinski and Weinstein (2000), and Martinek (1980) bivariate correlations (Pearson and Spearman) were assessed to investigate rank order stability. In correspondence with Kuklinski and Weinstein (2000), changes in the level of teacher expectation ratings were also assessed to get a first impression of teacher expectation stability at the within-student level. Moreover, we estimated empty three-level multilevel models using MLwiN 3 software (Charlton, Rasbash, Browne, Healy, & Cameron, 2020) to assess how the variances in teacher expectations and student achievement were associated at the within-student level (1) as well as the between-student (2) and class-level (3). From these models, intraclass correlation coefficients, ICCs, the correlation between two randomly drawn units (e.g., students) from the same higher-level unit (i.e., class) were determined (Snijders & Bosker, 2012).

Second, to be able to distinguish between absolute and relative stability, both a CLPM and a RI-CLPM were estimated. The CLPM most closely reflected methods of previous studies in which the rank order correlations were the main method for studying teacher expectation stability. The RI-CLPM is estimated to explicitly distinguish the between-person stability from the within-person stability. The most notable methodological difference between CLPM and RI-CLPM is the estimation of latent factors which partial out the stable between-person differences (i.e., stable between student differences in teacher expectations and mathematics achievement) in the manifest indicators (Becht et al., 2017; Burns et al., 2018; Hamaker et al., 2015). Due to these latent factors, the interpretation of the structural parameters in the RI-CLPM are also different from those of the CLPM. In the CLPM, the autoregressive parameters reflect the rank order stability of persons from one measurement occasion to the next. For example, a positive autoregressive parameter in the CLPM indicates that if a student is ranked highly on teacher expectations at the first measurement occasion the student is likely to be ranked highly as well on the next measurement occasion. In

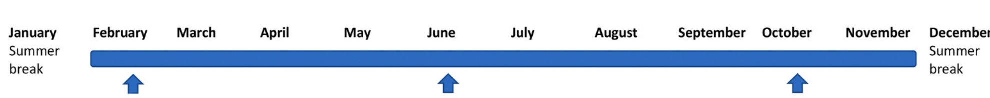

Fig. 1. Overview of the New Zealand primary school year and the measurements of teacher expectations (TE) and student mathematics achievement (Ach).

Table 2
Descriptive statistics of teachers' expectations and students' mathematics achievement.

	Occasion	N	Min.	Max.	Mean	SD	Skewness	Kurtosis
Teacher expectations	1	2536	1	7	4.66	1.425	−0.311	−0.201
	2	2536	1	7	4.52	1.449	−0.257	−0.278
	3	2536	1	7	4.54	1.404	−0.267	−0.244
Achievement	1	1907	−317.00	253.00	−8.29	78.00	−0.195	0.228
	2	1886	−293.00	252.00	1.57	80.23	0.073	−0.018
	3	1533	−313.00	352.00	9.66	87.17	0.033	0.219

Note. Occasion 1 refers to the first month of school year (February). Occasion 2 is middle of the school year (June). Occasion 3 is near the end of the school year (October).

contrast, for the RI-CLPM, the autoregressive parameters reflect the amount of within-person carry-over effect (e.g., Hamaker, 2012; Hamaker et al., 2015). That is, a positive autoregressive parameter in the RI-CLPM reflects the likelihood that occasions on which a person scored above his/her expected score are likely to be followed by occasions on which he/she still scores above the expected score again, and vice versa. For example, a positive autoregressive parameter in the RI-CLPM indicates that if a teacher's expectation for a particular student at the second measurement occasion was higher than expected given the previous level of expectations and student performance, then the teacher's expectation for this student at the third measurement occasion tended to be higher than the expected value as well. The interpretation of the cross-lagged parameters also changes due to the latent factors. In the CLPM, the cross-lagged parameters are usually interpreted as predicting change, that is, the extent to which change in one variable (i.e., teacher expectations) can be predicted from the individual's preceding deviation from the group mean on the other variable (i.e., mathematics achievement). In the RI-CLPM, they now reflect whether change from an individual's expected score on one variable (i.e., teacher expectations) is predicted from preceding deviations from the student's own mean on a second variable (i.e., mathematics achievement) while controlling for the structural change in y and the prior deviation from one's expected score on y.

Both models were estimated in a structural equation framework in MPlus v7 (Muthén & Muthén, 1998–2012) using the MLR estimator. Therefore, participants with any missing or non-response observations were retained in the analysis and the MLR estimation adjusted the likelihood function so that each case contributed information on the variables that were observed. Of all values in the dataset, 85.0% were observed. The pattern of missing values could best be described as multivariate and non-monotone, as missing values occurred in every mathematics test and missing values at the first occasions did not imply that students did not return for later measurements. Because the teacher expectation variables consisted of seven categories and skewness and kurtosis for the teacher expectation variables were close to zero, we considered this variable as a continuous variable. According to several researchers this can be done without harm to the analyses (e.g., Norman, 2010; Sullivan & Artino, 2013; Zumbo & Zimmerman, 1993).

Comparisons between the models were made by comparing a range of goodness-of-fit indices. For each model, chi-square values (Kline, 1998), Akaike (AIC) and Bayesian information criteria (BIC), the comparative fit index (CFI; > 0.95; Hu & Bentler, 1999), the Tucker–Lewis index (TLI; > 0.90; Hu & Bentler, 1999), standardized root mean square residual (SRMR, < 0.08; Hu & Bentler, 1999), root-mean-square error of approximation (RMSEA; < 0.06; Hu & Bentler, 1999) are reported. Given the size of the sample (2536 students) and the small number of degrees of freedom in both models, chi-square values and the root-mean-square error of approximation were considered suboptimal measures of model fit (Asparouhov & Muthén, 2018; Bergh, 2015).

3. Results

3.1. Replicating methods from previous teacher expectation stability studies

An overview of the bivariate correlations between teacher expectations and student achievement over the three measurement occasions is presented in Table 3. The teachers' expectations of students were highly correlated over the three measurement occasions, ranging from r = 0.736 to 0.946. Smaller, yet moderate to strong correlations were found between student achievement at the three measurement occasions, ranging from r = 0.349 to 0.820. Similarly, relatively moderate to strong associations were observed between the teachers' expectations and student performance within a particular measurement occasion, ranging from r = 0.318 to 0.614.

A contingency table of ratings of the teachers' expectations for mathematics between occasions 1 and 2 and between occasions 2 and 3 are presented in Table 4. Over half the students in the sample (53.1%) were rated similarly on the teacher expectation scale between the first measurement at the start of the year (February) and the second measurement halfway through the year (June). When change occurred, it was most commonly a one-level change (32.2%), however, changes were present up to five levels (potential scale 0–6). Changes occurred more frequently in the negative direction (29.9%), compared to positive changes (20.0%). Between the second (June) and the third measurement occasion towards the end of the year (October), expectation ratings did not change for 79.6% of the students, and if changes in rating occurred, they were 2 levels at the most. Between the second and third measurement occasion positive changes in teacher expectation ratings were slightly more frequent (11.6%) compared to negative changes (8.9%).

The results of empty multilevel (three-level) models are presented in Table 5. It seems that the percentage of variance at the within-student level was much larger for mathematics achievement (49.2%) in comparison to teacher expectations for individual students (16.6%). As a result, the ICCs at the student and teacher level are higher for teacher expectations (student level: 0.834, teacher level: 0.246) than for student mathematics achievement (student level: 0.508, teacher level: 0.185). This implies that teacher expectations are more strongly related to time-invariant person-specific characteristics of the student and the teacher compared to student achievement.

3.2. Teacher expectation stability regarding mathematics

Standardized model parameters (*stdyx*), standard errors and associated 95% confidence intervals for the CLPM and RI-CLPM for mathematics expectations and achievement are presented in Table 6. Additionally, the results of the CLPM and RI-CLPM are graphically presented in Figs. 2 and 3.

Regarding the CLPM, assuming teacher expectation stability solely as a between-person phenomenon, poor model fit was found; χ^2 = 1243.958, df = 4, p < .001, AIC = 78236.343, BIC = 78370.624, CFI = 0.891, TLI = 0.619, SRMR = 0.028, RMSEA = 0.350. The autoregressive paths for both the teacher expectations and students' mathematics achievement were substantial and statistically significant in the CLPM,

Table 3
Bivariate correlations between teacher expectations and student achievement for the three measurement occasions in mathematics.

		Teacher expectations			Mathematics achievement		
Pearson correlations	Measurement occasion	1	2	3	1	2	3
Teacher expectations	1	1					
	2	.736**	1				
	3	.837**	.941**	1			
Mathematics achievement	1	.524**	.623**	.638**	1		
	2	.503**	.614**	.625**	.820**	1	
	3	.236**	.319**	.318**	.349**	.357**	1
Spearman rankorder correlations	Measurement occasion	1	2	3	1	2	3
Teacher expectations	1	1					
	2	.733**	1				
	3	.831**	.938**	1			
Mathematics achievement	1	.531**	.619**	.639**	1		
	2	.502**	.609**	.623**	.825**	1	
	3	.243**	.317**	.320**	.363**	.364**	1

Note. ** $p < .001$. Measurement occasion 1 refers to the first month of school year (February). Measurement occasion 2 is middle of the school year (June). Measurement occasion 3 is near the end of the school year (October).

Table 4
Descriptive indices of teacher expectation stability.

	Teacher expectation rating measurement occasion 2						
Teacher expectation rating measurement occasion 1	1	2	3	4	5	6	7
1. Very much below average	35	10	8	5	0	0	0
2. Moderately below average	9	63	46	24	3	0	0
3. Just below average	15	19	112	70	14	2	0
4. Average	13	36	102	407	123	40	7
5. Just above average	6	14	35	161	318	91	15
6. Moderately above average	1	3	6	65	81	258	48
7. Very much above average	0	2	0	18	40	57	154
	Teacher expectation rating measurement occasion 3						
Teacher expectation rating measurement occasion 2	1	2	3	4	5	6	7
1. Very much below average	58	18	3	0	0	0	0
2. Moderately below average	5	95	40	7	0	0	0
3. Just below average	0	22	233	53	1	0	0
4. Average	0	3	39	602	98	8	0
5. Just above average	0	0	3	58	473	45	0
6. Moderately above average	0	0	0	3	51	374	20
7. Very much above average	0	0	0	0	4	37	183

from 0.287 ($SE = 0.030$, $p < .001$) for mathematics achievement between measurement occasions 2 and 3 to 0.884 ($SE = 0.007$, $p < .001$) and for teacher expectations in mathematics between measurement occasions 2 and 3. Regarding the cross-lagged paths, there was no systematic directional dominance in the CLPM. Between the first and second measurement occasions the cross-lagged path from achievement to expectations (β = 0.318, SE = 0.016, $p < .001$) was the stronger path, whereas between the second and third measurements, the cross-lagged path from expectations to achievement was the stronger one (β = 0.151, SE = 0.030, $p < .001$).

In contrast, the RI-CLPM, assuming that teacher expectation stability may be partly a between-student phenomenon and partly a within-student phenomenon, generally showed relatively good model fit as cut-off values for most model fit indices were met; $\chi^2 = 39.327$, $df = 1$, $p < .001$, AIC = 77037.711, BIC = 77189.508, CFI = 0.997, TLI = 0.952, SRMR = 0.029, RMSEA = 0.123. A Chi-square differences test for comparing model fit between CLPM and RI-CLPM, as they are nested models (Hamaker et al., 2015), clearly indicated that the RI-CLPM fitted the data better; $\Delta\chi^2 = 1204.631$, $df = 3$, $p < .001$.

At the within-student level, the substantive magnitude of these autoregressive parameters of the RI-CLPM was reduced compared to the CLPM. In all cases, the autoregressive parameters were still statistically significant. The reduction in the magnitude of the autoregressive parameters was stronger for teacher expectations compared to mathematics achievement. In the case of teacher expectations between measurement occasions 1 and 2, even the direction of the parameter changed; CLPM = 0.565 (SE = 0.014, $p < .001$), RI-CLPM = −0.380 (SE = 0.074, $p < .001$). This negative parameter implied that if a teacher at the first measurement occasion had an expectation that was relatively

Table 5
Results from three-level empty regression models for teacher expectations and mathematics achievement.

	Teacher Expectations				Math Achievement			
	b	SE_b	VPC[a]	ICC[b]	B	SE_b	VPC	ICC
Fixed Part								
Constant	4.581	0.078			−1.090	3.924		
Random Part								
Level: Teachers	0.498	0.082	0.246	0.246	1230.322	205.745	0.185	0.185
Level: Students	1.188	0.037	0.588	0.834	2154.105	114.488	0.324	0.508
Level: Measurements	0.336	0.007	0.166		3271.356	82.725	0.492	
Units: Teachers	89				90			
Units: Students	2536				2274			
Units: Measurements	7608				5359			
−2*loglikelihood:	19734.013				60862.894			

[a] VPC = Variance partitioning components (proportion of the total variance located at a particular level).
[b] ICC = Intra class correlation (ICC at the teacher level is the correlation between two randomly drawn students from the same class; ICC at the student level is the correlation between two randomly drawn measurements from the same student in the same class).

Table 6
Standardized (stdyx) results for the CLPM and RI-CLPM for mathematics expectations and achievement.

	CLPM				RI-CLPM			
	β	SE	P	95% CI	β	SE	p	95% CI
Correlations								
T1	0.540	0.016	<.001	[0.509, 0.571]	0.388	0.065	<.001	[0.261, 0.515]
Between-person					0.779	0.081	<.001	[0.620, 0.938]
Cross-lagged paths								
TE T1 → Ach T2	0.133	0.016	<.001	[0.102, 0.164]	0.256	0.031	<.001	[0.195, 0.317]
TE T2 → Ach T3	0.151	0.030	<.001	[0.092, 0.210]	0.100	0.040	0.013	[0.022, 0.178]
Ach T1 → TE T2	0.318	0.016	<.001	[0.287, 0.349]	0.733	0.056	<.001	[0.623, 0.843]
Ach T2 → TE T3	0.091	0.009	<.001	[0.073, 0.109]	0.542	0.045	<.001	[0.454, 0.630]
Autoregressive paths								
TE T1 → TE T2	0.565	0.014	<.001	[0.538, 0.592]	−0.380	0.074	<.001	[-0.525, −0.235]
TE T2 → TE T3	0.884	0.007	<.001	[0.870, 0.898]	0.365	0.045	<.001	[0.277, 0.453]
Ach T1 → Ach T2	0.760	0.012	<.001	[0.737, 0.784]	0.683	0.024	<.001	[0.636, 0.730]
Ach T2 → Ach T3	0.287	0.030	<.001	[0.228, 0.346]	0.184	0.079	0.020	[0.029, 0.339]
Residual correlations								
T2	0.153	0.023	<.001	[0.108, 0.198]	0.793	0.061	<.001	[0.673, 0.913]
T3	0.015	0.025	0.543	[-0.034, 0.064]	0.035	0.028	0.213	[-0.020, 0.090]

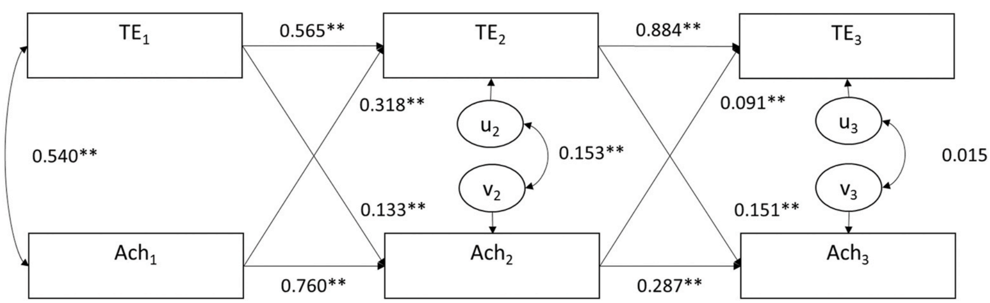

Fig. 2. Cross-lagged panel model for mathematics achievement and mathematics teacher expectations.
Note. **p* < .05, ***p* < .001. TE = Teacher Expectation, Ach = mathematics achievement. Subscripts refer to measurement occasions.

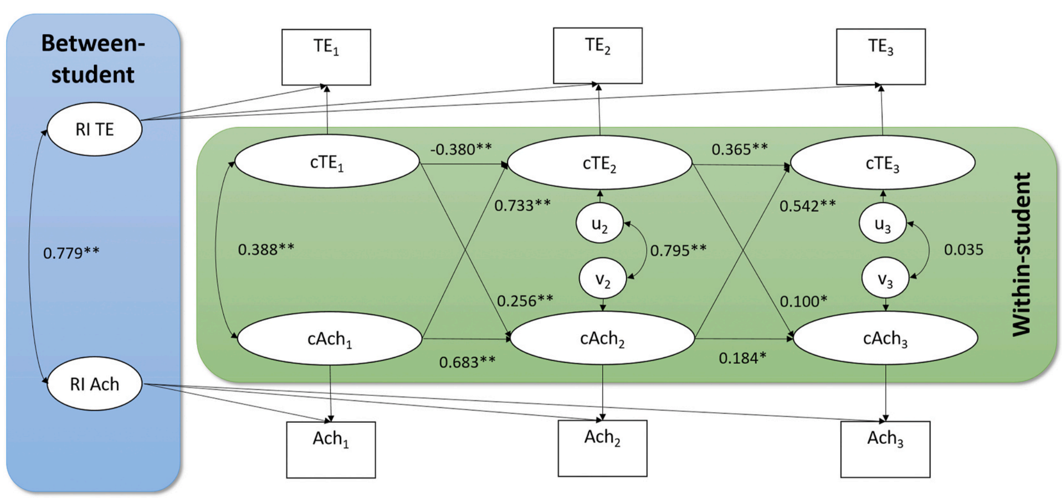

Fig. 3. Random intercept cross-lagged panel model for mathematics achievement and mathematics teacher expectations. *Note.* **p* < .05, ***p* < .001. TE = Teacher Expectation, Ach = mathematics achievement, RI = Random Intercept. Subscripts refer to measurement occasions.

high, the expectation at the second measurement occasion tended to be lower than the expected value based on previous performance. The cross-lagged paths (i.e., from teacher expectations at t to student achievement at $t+1$ and vice versa) at the within-student level also showed very different results over the two models, both in the substantive magnitude of the parameters as well as the directional dominance. In the RI-CLPM, both paths from student achievement to teacher expectations (i.e., adjustment paths) were clearly dominant over the paths from teacher expectations to student achievement (i.e., teacher expectation effect paths). Furthermore, in this model, the paths between the first and second measurement occasions were stronger compared to the paths between the second and third measurement occasions.

Finally, at the between-student level of the RI-CLPM, the random intercepts of teacher expectations and student achievement were highly correlated, indicating that the stable between-person differences in expectations were highly correlated with the stable between-person differences in mathematics achievement.

4. Discussion

The goal of the current study was to investigate the stability of teacher expectations of mathematics achievement for 2536 students in New Zealand elementary and middle schools within the time frame of a single school year (February, June, and October). Whereas some researchers have contended that teachers adjust the expectations for their students to fall in line with achievement (Brophy, 1983), indicating that expectations are at least to some degree unstable, others have noted that people are likely to adhere to their initial expectations (Babad, 2009; Kuklinski & Weinstein, 2000; Rubie-Davies et al., 2018), therefore suggesting stability in expectations. We aimed to study teacher expectation stability with a particular focus on within-student variability to be more in line with the idea that instability in expectations refers to teachers adjusting their expectations for individual students.

4.1. Replicating previous research

As a first step, we applied similar strategies to assess teacher expectation stability as in previous empirical studies regarding teacher expectation stability (e.g., Clifton, 1981; Good et al., 1980; Kuklinski & Weinstein, 2000; Martinek, 1980). Although this study was conducted several decades later, based on a New Zealand sample, and with different time lags compared to three of these previous studies in primary education, the results were quite similar. Teachers' expectations were highly correlated over the three measurement occasions within the school year, ranging from $r = 0.736$ to 0.946. This finding indicated that the rank order of students based on the teachers' expectations remained quite similar throughout the school year (Rubie-Davies et al., 2018).

Furthermore, the results regarding changes in the teacher expectation ratings strongly resembled the findings of Kuklinski and Weinstein (2000). Changes in teacher expectation ratings between the start (February) and the middle of the school year (June) were apparent for half the student sample. In the second half of the year (June–October), the prevalence of changes in expectations was much smaller. Similarly to Kuklinski and Weinstein (2000) and Wang et al. (2020), these results showed that teachers did adapt their expectations for some students and that expectations need to be considered unstable, at least to some degree. More prominent than in the study of Wang et al. (2020), instability occurred mainly during the first three or four months of the school year.

4.2. Testing stability using random intercept cross-lagged panel models

The random intercept cross-lagged panel model (RI-CLPM; Hamaker et al., 2015) fitted the data much better than the familiar cross-lagged panel model (CLPM; Rogosa, 1980), indicating that separating the within-student processes from stable between-student differences in teacher expectations and student performance indeed resulted in a better representation of the data. Furthermore, by partialling out the stable component of teacher expectations in the RI-CLPM, it became possible to demonstrate teacher expectation stability at the within student level. Finding smaller autoregressive coefficients in the RI-CLPM is consistent with previous research showing that the causal paths from the CLPM tend to be over-estimated as they fail to discriminate between- and within-person variation (Burns et al., 2019). A key finding of the current study is that teacher expectation stability is, consistent with theoretical notions, at least partly a phenomenon at the within-student level and therefore it should be investigated at the correct level of inference.

A second key finding of the current study is that, at the within-student level, both paths from student achievement to teacher expectations (i.e., adjustment paths) were stronger than the paths from teacher expectations to student achievement (i.e., teacher expectation effect paths). The magnitude of these teacher expectation effect paths was rather modest in this study. These findings were inconsistent with the claim that "[T]eachers' expectancies influence students' academic performance to a greater degree than students' performance influences teachers' expectancies" (Miller & Turnbull, 1986, p. 236). The significant and moderate to strong adjustment paths indicated that teachers did consider previous student performance and adapted their initial expectations for a student. More importantly, teachers continued to do so over the course of the school year. This latter finding is perhaps the most important contribution of this study to the existing literature in the domain of teacher expectations. Many studies have indicated that expectations are rather accurate (e.g., Brophy, 1983; Jussim & Eccles, 1992; Jussim & Harber, 2005; Südkamp, Kaiser, & Möller, 2012), which implies that teachers' expectations closely reflect the performance of students (e.g., Timmermans et al., 2015). However, these were almost exclusively studies with single teacher expectation measurements that were taken early in the academic year (Rubie-Davies et al., 2018), therefore, studies such as this one are important to provide empirical evidence that teachers keep on adjusting expectations throughout the school year.

Moreover, the cross-lagged paths between the first and second measurement occasions appeared stronger compared with the paths between the second and third measurement occasions. These outcomes demonstrate the hypothesized "Behavioral reference mode for reciprocal self-fulfilling prophecy" pattern of Levine and Doyle (2002). As teachers adjust their expectations and at the same time expectations influence subsequent student performance, the gap between performance and expectations is closed at a decreasing rate; fast at first and slower as the expectations become more in line with performance and vice versa. The accuracy of teachers' expectations increases when teachers have the chance to collect more information about their students (Dusek & Joseph, 1985), and the more accurate expectations become, the less is the possibility for self-fulfilling prophecies to arise and to continue to influence subsequent student performance.

Previous studies have indicated that at the between-student (e.g., Clifton, 1981; Good et al., 1980; Kuklinski & Weinstein, 2000; Martinek, 1980) or between-teacher levels (Rubie-Davies et al., 2018), stability has been found to be quite high. Although it may seem contradictory, the current study indicated that, consistent with theoretical notions, teacher expectations for individual students changed from one time point to the next as a function of student achievement. However, differences in the degree of stability at different levels may occur simultaneously. Even if teachers seem able to adjust their expectations for individual students upwards or downwards over the course of the year for individual students, this does not necessarily lead to large changes in the rank order of students. There may remain students for whom teachers have relatively high or low expectations and teachers who have relatively high or low expectations for all students in their classes (e.g., Rubie-Davies et al., 2018). A somewhat unexpected finding was the change in the autoregressive parameter of teacher expectations between measurement occasions one and two. This finding indicated that if a teacher at the first measurement occasion had an expectation that was relatively high compared to the average expectation over all measurement occasions for a particular student, the teacher's expectation at the second measurement occasion tended to be lower and vice versa.

4.3. Considerations regarding the study design

In the current study, however, we only investigated whether teachers adjusted their expectations based on students' previous e-asTTle mathematics test scores. Although e-asTTle has formative as well as summative evaluation capacity, not all teachers use it for formative evaluation. Many give a test early in the year to find out student levels but do not use it more formatively. Within e-asTTle, there are different reports available (individual student, class level, school level). If a teacher downloads the individual student report, it lists which objectives students understand and which ones they still need to learn. For the ones that they still need to learn, the teacher can click on any of the objectives, and it takes them to a page that provides resources for teachers to teach that particular concept. So, although e-asTTle is a standardized assessment, it can be equally used as a formative assessment tool.

It may well be the case that teachers also adapt their expectations for individual students based on other time-varying measures of student performance or behavior. Cross-sectional studies, for example, have indicated that student behaviors such as work habits, engagement, achievement motivation, and self-confidence predict teacher expectations above and beyond student performance (e.g., De Boer et al., 2010; Fitzpatrick, Côté-Lussier, & Blair, 2016; Kaiser, Retelsdorf, Südkamp, & Möller, 2013; Timmermans et al., 2016).

The current study, as with the few rare studies available to date that have taken into account multiple measurements of teacher expectations (e.g., Kuklinski & Weinstein, 2000; Martinek, 1980; Rubie-Davies et al., 2018), may have suffered from very long time lags; four months between measurements in the current study and eight weeks to six months in previous research. Teacher expectations are probably formed in the first weeks of the school year and are likely to become more stable during the course of the year, either because of corrections by teachers or because of self-fulfilling prophecy effects. The meta-analysis by Raudenbush (1984) showed that experimental studies inducing high teacher expectations only showed effects if the expectation was induced when the teacher had not met the students at all or within the first two weeks of the school year. Potential adjustments of expectations by teachers in the first weeks of the school year may not have been detected due to the four-month time lags in the current study. Optimal time lags (Dormann & Griffin, 2015) should be considered within the broader question of "when events occur, when they change, and how quickly they change" (Mitchell & James, 2001, p. 533). In the case of teacher expectations, this may imply that we need to investigate methods that enable more frequent measurements using smaller time lags in between while preventing memory effects (i.e., remembering answers provided to the expectation questions at previous measurements) to cause an overestimation of teacher expectation stability.

The current study made use of a single item to measure the teachers' expectations. As teachers were asked to provide their expectation for each individual student in their classes three times within the course of a year, using a single item was the most feasible to achieve a sample of sufficient size at both teacher- and student-level. However, an important prerequisite of any longitudinal study is that the same construct is analyzed over multiple measurement occasions, which is not guaranteed by the mere application of the same item, nor the application of the same scale (Seddig & Leitgöb, 2018). Response characteristics to survey questions may be different over time which implies that understanding has changed (e.g., Ferrer, Balluerka, & Widaman, 2008; Little, 2013; Stoel, van den Wittenboer, & Hox, 2004; Widaman, Ferrer, & Conger, 2010). In this particular case, it would be important to check whether the teachers' responses to the teacher expectation question were sufficiently similar at the very beginning of the school year when the teacher was just getting to know their students and later on in the school year when the teacher had become familiar with the students. This was not possible to assess because of the use of a single item; also, psychometric properties of teacher expectations could not be established. Future research within the field of teacher expectation may benefit from validation of the one-item scale (see Paas, Tuovinen, Tabbers, & Van Gerven, 2010 for an example on cognitive load), using a short standardized scale, or finding alternative novel modes of data collection to make it more feasible to ask teachers to answer multiple questions for each individual student, and testing whether the assumption of measurement invariance holds in a longitudinal teacher expectation study.

In this study, we only analyzed teacher expectation stability within the domain of mathematics. It is not uncommon in teacher expectation research to study expectations in a particular domain and to generalize to teacher expectations as a general construct (e.g., Kuklinski & Weinstein, 2000; Rubie-Davies, 2007). However, to the best of our knowledge it remains unclear whether expectations behave similarly over different subject domains. In regular mathematics tests it is usually clear whether a task was answered correctly or not by a student. Therefore, the assessment of and information on student performance is relatively straightforward in comparison with other domains. In domains such as writing or art, assessment may be harder and personal preferences of teachers may play a larger role. Also, teachers may be more likely (or it might be easier for teachers) to adjust their expectations based on clear and objective student achievement information in mathematics. Thus, it may be the case that expectations are more stable in domains in which tasks are more difficult to assess due to stable personal preferences or when information on student performance is more ambiguous. To investigate this hypothesis, future research may benefit from extending this paper to a variety of domains including those that are more or less easily assessed by teachers.

The recently developed RI-CLPM (Hamaker et al., 2015) applied in the current study showed superior fit to the data in comparison to the well-known CLPM (Rogosa, 1980), just as it has in other studies in the educational context (Burns et al., 2019; Ruzek & Schenke, 2019). Extensions of the RI-CLPM, of which some still need to be developed, may offer valuable opportunities to study subsequent research questions regarding teacher expectations. Examples are the possibility to include more than two indicator variables (e.g., multiple sources of information that teachers may use on which to build expectations), to examine between-group differences in stability (e.g., potential differences in stability between stigmatized vs. non-stigmatized groups), and most notably the extension to assess stability on more than two hierarchical levels (e.g., stability at the within-student level as well as at the level of between-students and between-classes). Investigating the previous extensions would require a panel study with at least three, but preferably more, measurements of teacher expectations for individual students over the course of a school year.

Nevertheless, this study included a large sample of students and teachers especially in comparison to previous empirical studies investigating teacher expectation stability (e.g., Good et al., 1980; Kuklinski & Weinstein, 2000; Martinek, 1980), and tracked teacher expectations at three moments across a full schooling year. Furthermore, it is the first study (to our knowledge) to explicitly examine teacher expectation stability at both the within-student and between-student level and to employ statistical methods more closely aligned to the theoretical considerations of adaptations in teacher expectations. Another strength of the paper is that it accounted for student achievement in examining stability, which has been a serious methodological flaw of most previous work on teacher expectation stability.

4.4. Practical implications

Many teacher training programs and textbooks stress the importance of having high expectations for students in a teacher's class. Moreover, creating awareness of potential biases in teacher expectations and the potential effects are common parts of interventions to raise teachers' expectations and subsequent student performance (De Boer, Timmermans, & Van Der Werf, 2018). Nevertheless, teacher expectation interventions are able to raise teachers' expectations as well as student performance, but it remains unclear the degree to which creating awareness is one predcitor of the positive outcomes of these interventions. For example, a teacher expectation intervention study by Rubie-Davies, Peterson, Sibley, and Rosenthal (2015) showed that when teachers raised their expectations of students and gave them more challenging work, the students showed increases in mathematics achievement the equivalent of a whole term's learning compared to students in the control group. This study provides a replicable model for teachers of the kinds of practices that need to accompany high expectations in order to raise student achievement. The results from the current study indicated that besides creating awareness of the importance of having high expectations, it may also be important to include a teacher expectation intervention that informs teachers and provides time to practice new behaviours related to how expectations of students are developed, that expectations of individual students may change during the course of a school year, and that it is important to regularly evaluate expectations based on achievement or other information gathered about students.

Creating awareness about the importance of having flexible expectations in teacher expectation interventions may include several topics. First, teachers need to become aware that the frequency of cases in which performance is consistent with expectations tends to get overestimated (Chapman, 1967), and confirming cases are more easily recalled (Crocker, 1981). Second, if new information becomes available showing that a student has more potential than a teacher originally believed, this should not be easily disregarded so that the student receives the optimal opportunity to learn. This may be a challenging task for teachers because of the processes by which teachers may disregard new performance information about a student. Newly obtained information may be considered ambiguous or teachers may attribute the new performance to factors outside the student (e.g., luck; Deaux & Emswiller, 1974; Regan et al., 1974). Processes such as these happen mostly at an unconscious level, therefore creating awareness is the first step necessary to promote change. Another complicating factor is that newly obtained information may come in multiple forms and from multiple sources. Performance information about students can come from standardized tests, like the ones included in this study, or from regular assignments, tasks or observations of behavior. Third, it is important to evaluate expectations on a regular basis, systematically schedule moments for recalibrating expectations and collect data on student performance, and carefully consider the capabilities of each individual student. Just creating awareness may not be sufficient to change teachers' behavior or their capacities for adjusting their expectations of individual students. Future research will need to show how teachers can be best supported in the process of adapting expectations to fall in line with newly obtained information.

Considering that high expectations have been associated with higher levels of achievement (e.g., Jussim & Harber, 2005; Timmermans et al., 2015), even if expectations are downgraded to be more in line with student achievement, the evidence suggests that higher expectations lead to higher-level opportunities to learn which, in turn, can increase student achievement (Rubie-Davies, 2015). Therefore, teachers need to be careful in terms of adopting lowered expectations. Students do not benefit from low expectations; but all students benefit when they are with teachers who have high expectations for all students.

Funding details

This study was supported by the Marsden Fund Council from Government funding administered by the Royal Society of New Zealand.

CRediT authorship contribution statement

Anneke C. Timmermans: Conceptualization, Methodology, Formal analysis, Writing – original draft, Visualization. **Christine M. Rubie-Davies:** Visualization, Resources, Data curation, Writing – review & editing, Funding acquisition. **Shengnan Wang:** Conceptualization, Writing – review & editing.

Declaration of competing interest

None.

References

Agirdag, O., Van Avermaet, P., & Van Houtte, M. (2013). School segregation and math achievement: A mixed-method study on the role of self-fulfilling prophecies. *Teachers College Record, 115*(3), 1–50.

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*, 731–746.

Asparouhov, T., & Muthén, B. (2018). *SRMR in mplus*. Retrieved from https://www.statmodel.com/download/SRMR2.pdf.

Babad, E. Y. (2009). *The social psychology of the classroom*. New York, NY: Routledge.

Becht, A. I., Nelemans, S. A., Dijk, M. P. A., Branje, S. J. T., Van Lier, P. A. C., Denissen, J. J. A., et al. (2017). Clear self, better relationships: Adolescents' self-concept clarity and relationship quality with parents and peers across 5 years. *Child Development, 88*, 1823–1833. https://doi.org/10.1111/cdev.12921

Bergh, D. (2015). Chi-squared test of fit and sample size: A comparison between a random sample approach and a chi-square value adjustment method. *Journal of Applied Measurement, 16*(2), 204–217.

Bohlmann, N. L., & Weinstein, R. S. (2013). Classroom context, teacher expectations, and cognitive level: Predicting children's math ability judgments. *Journal of Applied Developmental Psychology, 34*, 288–298. https://doi.org/10.1016/j.appdev.2013.06.003

Brophy, J. E. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 76*, 236–247. https://doi.org/10.1037/0022-0663.75.5.631

Brophy, J., & Good, T. (1970). Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology, 61*, 365–374.

Brophy, J. E., & Good, T. L. (1974). *Teacher-student relationships: Causes and consequences*. New York: Holt, Rinehart & Winston.

Burns, R. A., Crisp, D. A., & Burns, R. B. (2019). Re-examining the reciprocal effects model of self-concept, self-efficacy, and academic achievement in a comparison of the Cross-Lagged Panel and Random-Intercept Cross-Lagged Panel frameworks. *British Journal of Educational Psychology*, 1–15.

Chalabaev, A., Sarrazin, P., Trouilloud, D., & Jussim, L. (2009). Can sex-undifferentiated teacher expectations mask an influence of sex stereotypes? Alternative forms of sex bias in teacher expectations. *Journal of Applied Social Psychology, 39*, 2469–2498. https://doi.org/10.1111/j.1559-1816.2009.00534.x

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior, 6*, 151–155.

Charlton, C., Rasbash, J., Browne, W. J., Healy, M., & Cameron, B. (2020). *MLwiN version 3.05*. Centre for Multilevel Modelling, University of Bristol.

Clifton, R. A. (1981). The effects of students' ethnicity and sex on the expectations of teachers. *Interchange, 12*, 31–38.

Crocker, C. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90*, 272–292.

Darley, J. M., & Fazio, R. H. (1980). Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist, 35*, 867–881.

De Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology, 102*, 168–179. https://doi.org/10.1037/a0017289

De Boer, H., Timmermans, A. C., & Van Der Werf, M. P. C. (2018). The effects of teacher expectation interventions on teachers' expectations and student achievement: Narrative review and meta-analysis. *Educational Research and Evaluation, 24*(3–5), 180–200. https://doi.org/10.1080/13803611.2018.1550834

Deaux, K., & Emswiller, T. (1974). Explanations of successful performance on sex-linked tasks: What is skill for the male is luck for the female. *Journal of Personality and Social Psychology, 29*, 80–85.

Dormann, C., & Griffin, M. A. (2015). Optimal time lags in panel studies. *Psychological Methods, 20*, 489–505.

Doyle, W. (1977). Learning the classroom environment: An ecological analysis. *Journal of Teacher Education, 28*(6), 51–55.

Dusek, J. B., & Joseph, G. (1985). The bases of teacher expectancies. In J. B. Dusek (Ed.), *Teacher expectancies* (pp. 229–250). Hillsdale, NY: Erlbaum.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology, 4*(1), 22–36.

Fitzpatrick, C., Côté-Lussier, C., & Blair, C. (2016). Dressed and groomed for success in elementary school: Student appearance and academic adjustment. *The Elementary School Journal, 117*, 30–45.

Glock, S., & Krolak-Schwerdt, S. (2013). Does nationality matter? The impact of stereotypical expectations of student teachers' judgments. *Social Psychology of Education, 16*, 111–127.

Good, T. L., Cooper, H. M., & Blakey, S. L. (1980). Classroom interaction as a function expectations, student sex, and time of year. *Journal of Educational Psychology, 72*, 378–385.

Good, T. L., & Lavigne, A. L. (2018). *Looking in classrooms* (11th ed.). New York, NY: Routledge.

Gustavsson, J. P., Weinryb, R. M., Goransson, S., Pedersen, N. L., & Asberg, M. (1997). Stability and predictive ability of personality traits across 9 years. *Personality and Individual Differences, 22*, 783–791.
Hamaker, E. L. (2012). Why researchers should think "within-person" a paradigmatic rationale. In M. R. Mehl, & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). New York, NY: Guilford Publications.
Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20*, 102–116. https://doi.org/10.1037/a0038889
Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology, 101*, 662–670. https://doi.org/10.1037/a0014306
Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118
Jussim, L. (1986). Self-fulfilling prophecies: A theoretical and integrative review. *Psychological Review, 93*, 429–445.
Jussim, L., & Eccles, J. S. (1992). Teacher expectations II: Construction and reflection of students achievement. *Journal of Personality and Social Psychology, 63*, 947–961. https://doi.org/10.1037/0022-3514.63.6.947
Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*, 131–155.
Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28*, 73–84. https://doi.org/10.1016/j.learninstruc.2013.06.001
Kelly, S., & Carbonaro, W. (2012). Curriculum tracking and teacher expectations: Evidence from discrepant course taking models. *Social Psychology of Education, 15*, 271–294. https://doi.org/10.1007/s11218-012-9182-6
Kline, R. B. (1998). *Principles and practice of structural equation modeling. Methodology in the social sciences*. New York, NY: Guilford Press.
Kuklinski, M. R., & Weinstein, R. S. (2000). Classroom and grade level differences in the stability of teacher expectations and perceived differential treatment. *Learning Environments Research, 3*, 1–34. https://doi.org/10.1023/A:1009904718353
Kuklinski, M. R., & Weinstein, R. S. (2001). Classroom and developmental differences in a path model of teacher expectancy effects. *Child Development, 72*, 1554–1578. https://doi.org/10.1111/1467-8624.00365
Levine, R. L., & Doyle, J. K. (2002). Modeling generic structures and patterns in social psychology. In P. I. Davidsen, E. Mollona, V. G. Diker, R. S. Langer, & J. I. Rowe (Eds.), *Proceedings of the 20th international conference of the system dynamics society*. Retrieved http://www.systemdynamics.org/conferences/2002/index.html.
Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.
Marks, G. (2016). Explaining the substantial inter-domain and over-time correlations in student achievement: The importance of stable student attributes. *Educational Research and Evaluation, 22*(1–2), 45–64. https://doi.org/10.1080/13803611.2016.1191359
Martinek, T. J. (1980). Stability of teachers' expectations for elementary school aged children. *Perceptual & Motor Skills, 51*(suppl2), 1269–1270.
McKown, C., & Weinstein, R. S. (2002). Modelling the role of child ethnicity and gender in children's differential response to teacher expectations. *Journal of Applied Social Psychology, 32*, 159–184.
McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology, 46*, 235–261. https://doi.org/10.1016/j.jsp.2007.05.001
Merton, R. K. (1948). The self-fulfilling prophecy. *Antioch Review, 8*, 193–210.
Miller, D. T., & Turnbull, W. (1986). Expectancies and interpersonal processes. In M. R. Rosenzweig, & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 37, pp. 233–256).
Mitchell, T. E., & James, L. R. (2001). Building better theory: Time and the specification of when things happen. *Academy of Management Review, 26*, 530–547.
Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y
Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. (2010). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63–71. https://doi.org/10.1207/S15326985EP3801_8
Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76*, 85–97.
Regan, D. T., Straus, E., & Fazio, R. (1974). Liking and the attribution process. *Journal of Experimental Social Psychology, 10*, 385–397.
Rogosa, D. (1980). A critique of cross-lagged correlation. *Psychological Bulletin, 88*(2), 245–258.
Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology, 76*, 1000–1027.
Rosenholtz, S. J., & Simpson, C. (1984). The formation of ability conceptions: Developmental trend or social construction? *Review of Educational Research, 54*, 31–63.
Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York, NY: Holt, Rinehart & Winston.
Rubie-Davies, C. M. (2007). Classroom interactions: Exploring the practices of high- and low-expectation teachers. *British Journal of Educational Psychology, 77*, 289–306. https://doi.org/10.1348/000709906X101601
Rubie-Davies, C. M. (2015a). *Becoming a high expectation teacher: Raising the Bar*. London: Routledge.
Rubie-Davies, C. M. (2015b). Expecting more: Teacher differences as moderators of expectancy effects. In C. M. Rubie-Davies, J. M. Stephens, & P. Watson (Eds.), *The Routledge international handbook of social psychology of the classroom*. London, UK: Routledge.
Rubie-Davies, C. M., Peterson, E. R., Sibley, C. G., & Rosenthal, R. (2015). A teacher expectation intervention: Modelling the practices of high expectation teachers. *Contemporary Educational Psychology, 40*, 72–85. https://doi.org/10.1016/j.cedpsych.2014.03.003
Rubie-Davies, C. M., Watson, P. W., Flint, A., Garrett, L., & McDonald, L. (2018). Viewing students consistently: How stable are teachers' expectations? *Educational Research and Evaluation, 24*(3–5), 221–240. https://doi.org/10.1080/13803611.2018.1550836
Rubie-Davies, C. M., Weinstein, R. S., Huang, F. L., Gregory, A., Cowan, P. A., & Cowan, C. P. (2014). Successive teacher expectation effects across the early school years. *Journal of Applied Developmental Psychology, 35*, 181–191. https://doi.org/10.1016/j.appdev.2014.03.006
Ruzek, E. A., & Schenke, K. (2019). The tenuous link between classroom perceptions and motivation: A within-person longitudinal study. *Journal of Educational Psychology, 111*(5), 903–917.
Seddig, D., & Leitgöb, H. (2018). Approximate measurement invariance and longitudinal confirmatory factor analysis: Concept and application with panel data. *Survey Research Methods, 12*, 29–43. https://doi.org/10.18148/srm/2018.v12i1.7210
Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research, 46*, 553–611.
Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.
Sorhagen, N. S. (2013). Early teacher expectations disproportionately affect poor children's high school performance. *Journal of Educational Psychology, 105*, 465–477. https://doi.org/10.1037/a0031754
Stoel, R. D., van den Wittenboer, G., & Hox, J. (2004). Methodological issues in the application of the latent growth curve model. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models: Theory and applications* (pp. 241–261). Amsterdam, the Netherlands: Kluwer Academic Publishers.
Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743–762. https://doi.org/10.1037/a0027627
Sullivan, G. M., & Artino, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education, 5*(4), 541–542. https://doi.org/10.4300/jgme-5-4-18
Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology, 99*, 689–705.
Timmermans, A. C., de Boer, H., & van der Werf, M. P. C. (2016). An investigation of the relationship between teachers' expectations and teachers' perceptions of student attributes. *Social Psychology of Education, 19*, 217–240. https://doi.org/10.1007/s11218-015-9326-6
Timmermans, A. C., Kuyper, H., & Van der Werf, M. P. C. (2015). Accurate, inaccurate or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology, 85*, 459–478.
Timmermans, A. C., & Rubie-Davies, C. M. (2018). Do teachers differ in the level of expectations or in the extent to which they differentiate in expectations? Relations between teacher-level expectations, teacher background and beliefs, and subsequent student performance. *Educational Research and Evaluation, 24*(3–5), 241–263. https://doi.org/10.1080/13803611.2018.1550837
Timmermans, A. C., & van der Werf, M. P. C. (2018). Accounting for previous performance of students by means of growth curves analyses to estimate the size, stability and consistency of school effects. *Educational Research and Evaluation, 23*(5–6), 221–246, 3803611.2017.1455300.
Timmermans, A. C., van der Werf, M. P. C., & Rubie-Davies, C. M. (2019). The interpersonal character of teacher expectations: The perceived teacher-student relationship as an antecedent of teachers' track recommendations. *Journal of School Psychology, 73*, 114–130. https://doi.org/10.1016/j.jsp.2019.02.004
Tyler, K. M., & Boelter, C. M. (2008). Linking black middle school students' perceptions of teachers' expectations to academic engagement and efficacy. *The Negro Educational Review, 59*(1–2), 27–44.
Wang, S., Rubie-Davies, C. M., & Meissel, K. (2020). The stability and trajectories of teacher expectations: Student achievement level as a moderator. *Learning and Individual Differences, 78*, 101819. https://doi.org/10.1016/j.lindif.2019.101819
Weinstein, R. S. (2002). *Reaching higher: The power of expectations in schooling*. Cambridge, MA: Harvard University Press.
Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4*(1), 10–18.
Zhu, M., & Urhahne, D. (2015). Teachers' judgments of students' foreign-language achievement. *European Journal of Psychology of Education, 30*, 21–39.
Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology/Psychologie Canadienne, 34*(4), 390–400. https://doi.org/10.1037/h0078865

Anneke Timmermans is an associate professor at the Faculty of Behavioural and Social Sciences of Groningen University. Her research interests are teacher expectations, factors related to the transition from primary to secondary schools in tracked educational systems, multilevel modeling, and the validity and reliability of school performance indicators. In

2016, Anneke received an Early Career Researcher Award from the British Journal of Educational Psychology for the paper "*Accurate, inaccurate, or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education?*".

Christine Rubie-Davies is a Professor of Education, Faculty of Education and Social Work, University of Auckland, New Zealand. Her primary research interests are teacher expectations and beliefs that moderate expectancy effects. Christine primarily focuses on teacher expectation effects for disadvantaged groups. She has published six books and numerous articles and chapters in prestigious presses, has won national and international awards for research, teaching and service, and is a Fellow of the American Psychological Association and the Association for Psychological Science.

Shengnan Wang is a research fellow in the Faculty of Education and Social Work at the University of Auckland, New Zealand. Her research focuses on teacher expectations and teacher expectation effects in the Chinese secondary school context. Shengnan is particularly interested in teachers' behaviors that communicate differential expectations to students, the stability of teacher expectations, and teacher expectations for marginalized and vulnerable groups of students in China.