Supplementary Material

1 Online Appendix S0: Data analysis with cumulative link models

As the dependent variable of this study was a single item (i.e., “How much would you like to sit next to this student?”), the preconditions of normally distributed error terms for multilevel models were not met (Gelman & Hill, 2006). To consider the scale of the dependent variable, we used ordinal multilevel models and tested our analysis with cumulative link models of the ordinal package (Christensen, 2015) in R (R Core Development Team, 2020). Compared with linear models, these models treat the data as ordinal, but do not contain assumptions about the distance between response categories.

Cumulative link models assume that the response variable is ordinal and that any observation Y_i falls within a category j = 1 to J categories (Christensen, 2015). For each observation i the probability how likely an observation falls into a category that is smaller or equal to j is estimated $\pi_{ij} = \mathrm{P}(Y_i \leq j)$. In a cumulative logit model a logit link is used to estimate these probabilities as a function of the explanatory variables: $logit_{(\pi_{ij})} = \theta_j - x_i^T \beta$. Thereby, the $\theta_j$parameters represent the intercepts of the model for each category (i.e., the thresholds between two categories), x_i is a vector of all the explanatory variables for the ith observation, and β represents a set of regression parameters. This means that the resulting β coefficients represent a linear combination of the estimated effects that is independent of j. In other words, the estimated effects are the same for each response category. This linear combination of the effects can be understood as a latent variable. As higher values reflect a smaller probability that a category falls within a category that is equal or smaller than j, higher values of this latent variable reflect a higher probability for choosing higher categories than j.

In the cumulative link model, thresholds between the different categories are estimated (represented by θ_j); thereby, it is possible to assume a model structure in which the distances between the thresholds are equidistant between the categories. To test which model structure fit the data best, we compared a model with equidistant thresholds to a model with freely estimated thresholds, whereby the model with freely estimated thresholds did not fit the data significantly better. Therefore, we used a model with equidistant thresholds.

Supplementary analyses S2: Drivers of ethnic homophily: Who is choosing whom

In addition to analyzing ethnic homophily, we were interested in finding out whether German students expressed higher in-group bias than non-German students. Therefore, we conducted separate exploratory analyses including the dyadic match by ego interaction terms to the model (e.g., do German pupils rate German classmates more positively relative to mixed dyads? And similarly, do non-German pupils rate non-German classmates more positively relative to mixed dyads?). For these analyses, we created a factorial variable that expressed if both students were either German, non-German, or if the dyads were mixed[1].

First, we examined whether students with German and non-German backgrounds had the same probability of expressing ethnic homophily (i.e., in-group bias). The results (see Table S2, Step 1) showed that students with a German background favored German students over interethnic ethnic dyads; in contrast, students with a non-German background (nGb) rated non-German students less positively relative to interethnic dyads. In addition, the results (see Table S2, step 3) indicated that ethnic in-group bias significantly depended on ethnic diversity. The interaction plot showed that students with a German background expressed more in-group bias in classrooms with high ethnic diversity compared to classrooms with low ethnic diversity (see Figure S1; $z = 6.03$ $p < .001$). In contrast, nGb students showed a lower preference for nGb students in classrooms with higher ethnic diversity as compared to classrooms with lower ethnic diversity. However, this difference was not statistically significant, according to the post-hoc tests, $z = 2.12$, $p = .276$.

Regarding faultlines, the results (see Table S2, step 3) suggested that the probability that students with a German background preferred German students depended on faultline strength. The interaction plot demonstrated that they expressed more in-group bias in classrooms with strong faultlines as compared to classrooms with weak faultlines (see Figure S2; $z = 3.10$, $p = .024$), whereas this was not the case for students with nGb, $z = 1.20$, $p = .838$.

Regarding teacher care, the results (see Table S2, step 3) showed a significant interaction effect for teacher care regarding how German students rated German students compared to interethnic dyads. When examining the interaction (see Figure S3), it became apparent that students with a German background still expressed in-group preference relative to interethnic-dyads when teacher care was high, $z = 12.22$, $p < .001$. However, this difference was smaller compared to when teacher care was low, $z = 14.97$, $p < .001$. Importantly, on average, students in interethnic dyads showed more positive peer ratings when teacher care was high as compared to when teacher care was low, $z = 4.02$, $p < .001$ (see Figure S3).

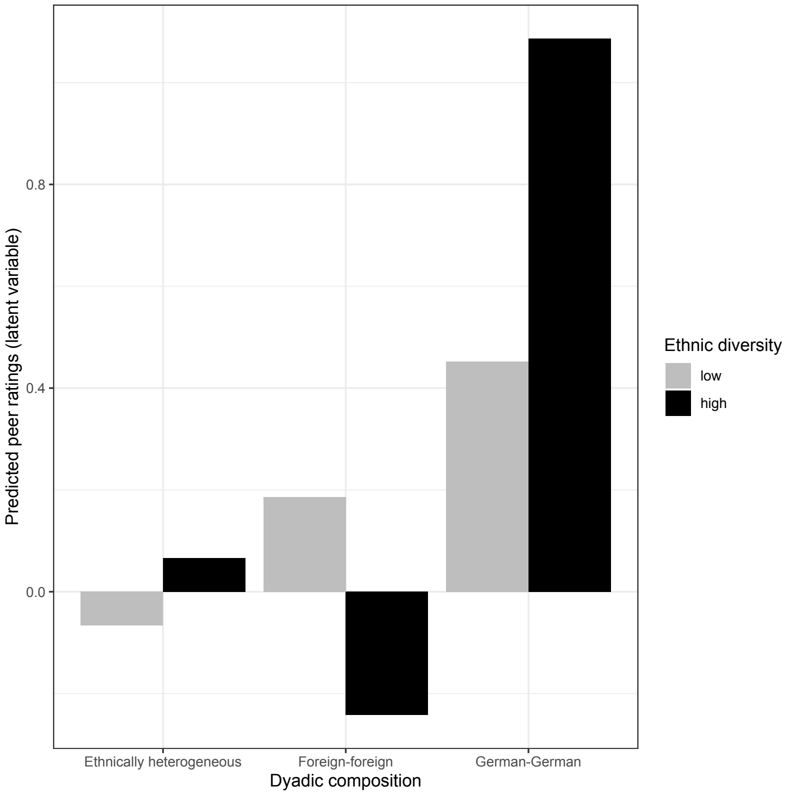
Lastly, the results from the significant three-way interactions between teacher care, faultlines, and dyadic composition are plotted in Figure S4. Similar to the results that did not differentiate between the ego*dyadic match terms, the post-hoc contrasts revealed that German students showed more in-group bias (i.e., relative to interethnic dyads) in classrooms with strong faultlines as compared to classrooms with weak faultlines when teacher care was low, $z = 3.20$, $p = .060$. This was not the case when teacher care was high, $z = 1.89$, $p = .764$. Importantly, in classrooms with strong faultlines, interethnic dyads were rated significantly more positive when teacher care was high as compared to when it was low, $z = 3.91$, $p = .005$ (see Figure S4).

[1] As this term “own ethnicity x dyadic match in ethnicity” represented a factor with three levels (both German, both non-German background, both different) that contained information about ego’s and alter’s ethnicity, we had to exclude alter’s ethnicity from the model for this specific analysis, as we would encounter problems of singularity otherwise.

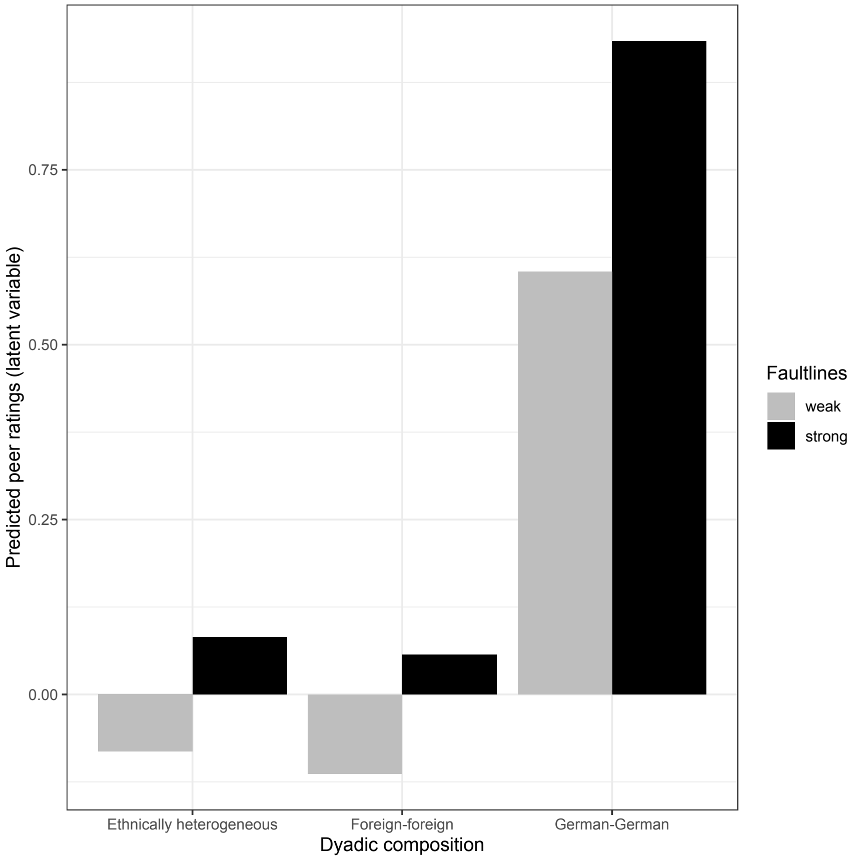
Table S2 Results of the multilevel cumulative link models predicting students' inclusion preferences (*N* = 23727 peer ratings by 1299 fourth graders in 70 classrooms)

	Step 1	Step 3	Step 4
	γ (*SE*)	γ (*SE*)	γ (*SE*)
Fixed Effects Level 2: Ego attributes			
Sex (0 = male)	0.09 (0.07)	0.10 (0.07)	0.10 (0.07)
Ethnicity (0 = German)	0.67 (0.08)***	0.63 (0.08)***	0.63 (0.08)***
Ego * dyadic match of alter and ego			
Dyadic: Both male	2.39 (0.04)***	2.38 (0.04)***	2.38 (0.04)***
Dyadic: Both female	1.98 (0.04)***	1.97 (0.04)***	1.97 (0.04)***
Dyadic: Both German background (Gb)	0.81 (0.04)***	0.77 (0.04)***	0.78 (0.04)***
Dyadic: Both non-German background (nGb)	-0.28 (0.06)***	-0.03 (0.09)	-0.04 (0.09)
Fixed Effects Level 3: Classroom attributes			
Gender diversity		1.17 (0.89)	1.18 (0.87)
Ethnic diversity		0.36 (0.31)	0.39 (0.32)
Faultlines		0.83 (0.52)	0.84 (0.53)
Teacher care		0.27 (0.07)***	0.33 (0.07)***
Teacher care * ethnic diversity			-0.73 (0.45)
Teacher care * faultlines			0.53 (0.71)
Ego * dyadic match and classroom attributes			
Dyadic: Ethnic diversity * both Gb		1.38 (0.26)***	1.35 (0.26)***
Dyadic: Ethnic diversity * both nGb		-1.54 (0.50)**	-1.46 (0.50)**
Dyadic: Faultlines * both Gb		0.85 (0.45)†	0.92 (0.45)*
Dyadic: Faultlines * both nGb		0.04 (0.61)	0.03 (0.61)
Dyadic: Teacher care * both Gb		-0.16 (0.07)*	-0.20 (0.07)**
Dyadic: Teacher care * both nGb		-0.02 (0.11)	-0.03 (0.16)
Dyadic: Teacher care * ethnic diversity*both Gb			0.79 (0.46)†
Dyadic: Teacher care * ethnic diversity*both nGb			-0.39 (0.90)
Dyadic: Teacher care * faultlines*both Gb			-1.39 (0.74)†
Dyadic: Teacher care * faultlines*both nGb			-2.22 (1.05)*
AIC	49699.43	47596.80	47593.23
Cond. H	110	8300	18000

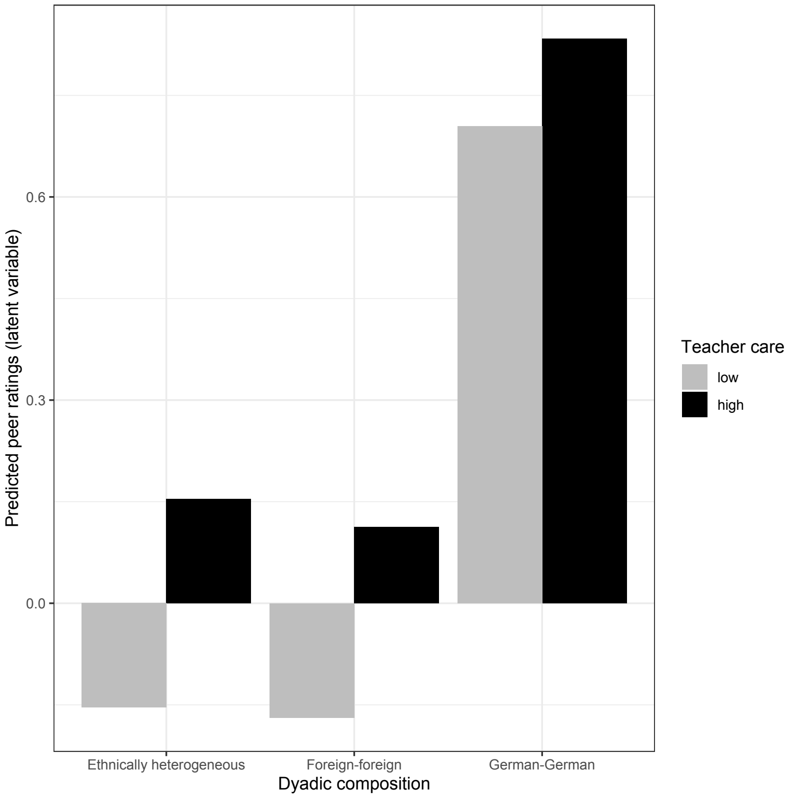
Note. The variables gender diversity, ethnic diversity, faultlines, and teacher care were mean-centered. † < .10, * *p* < .05, ** *p* < .01, *** *p* < .001, two-tailed.



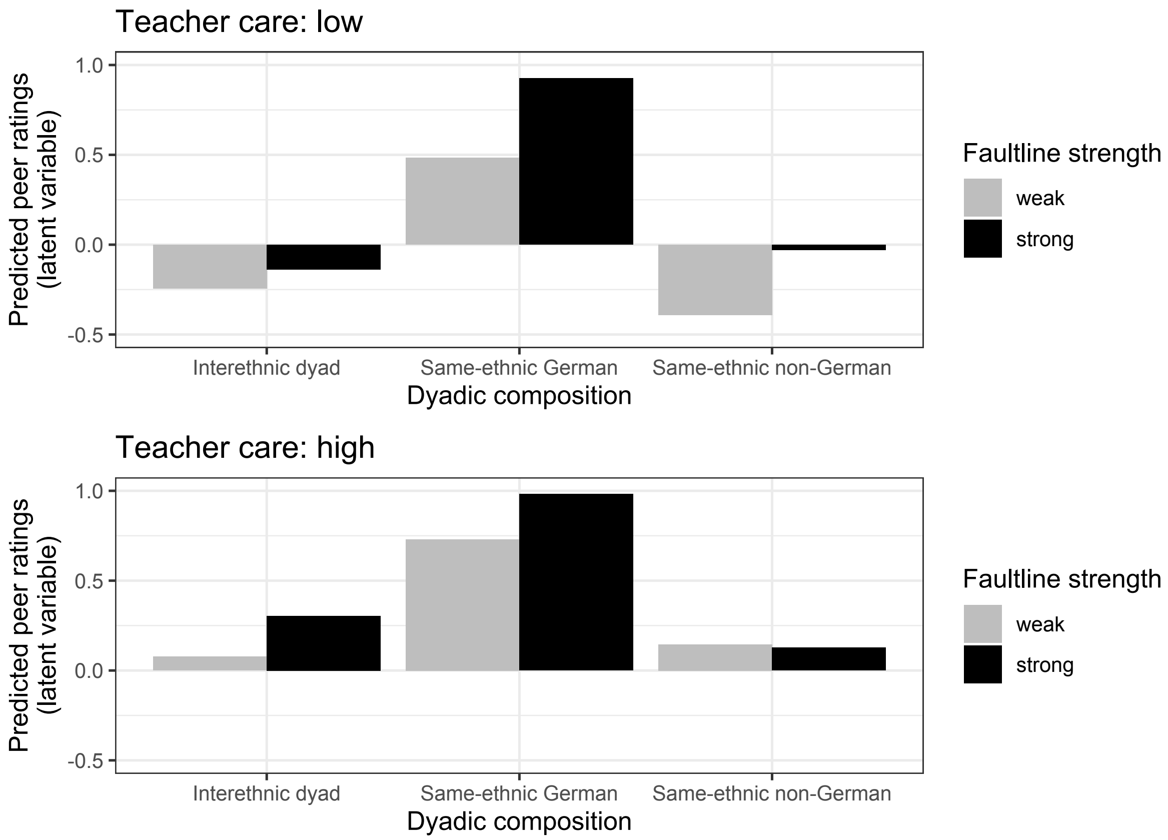
Supplementary Figure S1. Students' ratings of their classmates ("How much would you like to sit next to X?") for interethnic, non-German – non-German background, and German – German background dyads and classrooms with low versus high ethnic diversity. The latent variable represents the probability for choosing higher categories.



Supplementary Figure S2. Students' ratings of their classmates ("How much would you like to sit next to X?") for interethnic, non-German – non-German background, and German – German background dyads and classrooms with weak versus strong faultlines. The latent variable represents the probability for choosing higher categories.



Supplementary Figure S3. Students' ratings of their classmates ("How much would you like to sit next to X?") for interethnic, non-German – non-German background, and German – German background dyads and classrooms with low versus high teacher care. The latent variable represents the probability for choosing higher categories.



Supplementary Figure S4. Students' ratings of their classmates ("How much would you like to sit next to X?") for interethnic, non-German – non-German background, and German – German background dyads, depending on faultline strength and teacher care. The latent variable represents the probability for choosing higher categories.