WORKING PAPER 5/2026

# Ask the Child or the Parent? Survey Design and the Measurement of School Satisfaction

**Daniela Andrén**

ISSN 1403-0586

Örebro University School of Business
SE-701 82 Örebro, Sweden

# Ask the Child or the Parent? Survey Design and the Measurement of School Satisfaction


Daniela Andrén

Örebro University, Sweden

daniela.andren@oru.se



**Abstract**

In 1965, about one thousand children in the third grade of a midsize Swedish town answered a short questionnaire about school, and their parents answered a parallel questionnaire about the same child (a design repeated in grade six and grade nine). This article uses these paired reports to revisit, in a new construct and country, two sensitivities highlighted by Conti and Pudney (2011): that response-scale format and labeling can reshape reported satisfaction distributions, and that response conditions can induce systematic, gender-differentiated departures from classical measurement error. Parents concentrate on the conventional positive category while children spread across the scale; the distributional distance between informants is largest in grade six, where both answered fully labeled five-point scales. An integrative latent model attributes to the child report extreme-category misclassification probabilities of 0.14–0.34 and to the parent report a systematic threshold shift (context/informant perspective) with noise standard deviation close to one (estimates close to the original British ones). The choice of informant changes econometric conclusions: conditional on achievement, cognitive ability is negatively associated with satisfaction only in the child's own reports, and classroom-context coefficients switch between informants. Measured satisfaction is partly an artifact of who is asked and how.



**Keywords:** survey design; school satisfaction; informant effects; ordinal response; measurement error.

**JEL codes:** C81, C25, I21, J16.

# 1 Introduction

Subjective satisfaction data now sit at the center of large empirical literatures in economics, yet the measurement process that produces them is rarely part of the econometric model (Bertrand and Mullainathan, 2001; Bound, Brown and Mathiowetz, 2001). Conti and Pudney (2011) showed, using two quasi-experiments in the British Household Panel Survey, that two apparently minor design features (the textual labeling of response categories and the context in which the question is answered) distort reported job satisfaction enough to overturn substantive conclusions, particularly about gender differences. Their warning has been influential, but it rests on one construct (job satisfaction), one survey, and one country.

This article revisits the logic of Conti and Pudney (2011) in a setting they could not observe: the same satisfaction construct reported simultaneously by two different informants. In a Swedish longitudinal cohort study that began in 1965, the pupils of an entire school-year cohort answered classroom questionnaires about how much they liked school, while their parents answered a home questionnaire containing a parallel question about the same child. The pairing was repeated three times (in grade three (age about ten), grade six (age about thirteen), and grade nine (age about sixteen)), and the two instruments vary in the two dimensions emphasized by the original study: response-scale format/labeling and the conditions of response (a supervised classroom versus the family home).

One caveat governs the interpretation throughout. In the British design, the same person answered the same question in two modes; here, two informants answer about the same target. The child–parent gap therefore blends genuine design and context effects with a real difference in perspective between child and parent. The analysis does not claim to separate these; it asks the question that matters for applied work: does the choice of instrument (equivalently, of informant) change the measured distribution of satisfaction and the econometric conclusions drawn from it? The answer, as in the original study, is yes, and in gender-differentiated ways. The paper's added value is (i) to document how large and structured the informant gap is when two reports are collected contemporaneously on the same target, (ii) to show that the discrepancy is systematically related to observables rather than behaving like classical noise, and (iii) to quantify how these measurement differences propagate into standard satisfaction regressions.

The analysis follows the original sequence of exercises: distributional comparisons summarized by Kullback–Leibler differentials with cluster-bootstrap inference; a multinomial logit model of the discrepancy between the two reports; random-effects ordered probit models fitted separately to each report series; and an integrative latent-factor model in which one report suffers extreme-category misclassification and the other a covariate-driven threshold shift (capturing both response conditions and informant perspective). Three results stand out. First, parents concentrate massively on the conventional positive category ("good"), while children use the extremes; the informant divergence is largest precisely where the two scales are most comparable. Second, the discrepancy between reports is not classical noise: it is associated with the child's school achievement, teacher-rated behavior, peer status, and classroom context, with different loadings for boys and girls. Third, the informant choice propagates into estimation: cognitive ability conditional on achievement carries a negative coefficient only in the child's own equation, and classroom variables switch sign between informants. The integrative model, estimated on the grade-six pair, returns misclassification and context-noise parameters remarkably close to the original British estimates.

The article proceeds as follows. Section 2 summarizes the original study. Section 3 describes the cohort study. Section 4 maps the original design onto the informant pairs and states the hypotheses. Section 5 details measures and samples. Section 6 reports the four blocks of results. Section 7 discusses feasibility and limitations, and Section 8 concludes.

# 2 The original study

Conti and Pudney (2011) exploit two design changes in the British Household Panel Survey. First, between 1991 and 1992 the show-card for the seven-point overall job satisfaction question changed from partial labeling (only categories one, four, and seven carried text) to full labeling of all seven categories. The 1991 distribution differs sharply from the stable 1992–93 distributions (Kullback–Leibler differentials of 0.105 for men and 0.069 for women against a within-design differential of 0.005), with the distortion concentrated on the three originally labeled categories and stronger for women, consistent with the well-documented female advantage in verbal orientation (Halpern, 2000). The Kullback–Leibler metric provides a single-number summary of global distributional reshaping across ordered categories, rather than focusing only on moments or a single cutpoint. Second, from 1996 the same construct was measured twice for the same respondents: orally, in a household interview that others could overhear, and privately, in a self-completion questionnaire with only the extreme categories labeled. Only forty-five percent of respondents gave the same answer twice. A multinomial logit of the discrepancy shows systematic, non-classical structure: married men and respondents whose partner was present under-reported satisfaction orally, high-wage men under-reported less, and the patterns differed by sex. Random-effects ordered probit models (static, Mundlak, and dynamic with Wooldridge (2005) initial conditions) fitted separately to the two measures showed that the well-known finding that women's job satisfaction does not respond to wages (Clark, 1997) appears only in the interview measure; in the private measure women's wage coefficient is positive and significant, dissolving the part-time work puzzle of Booth and van Ours (2008). An integrative latent model attributed to the self-completion measure extreme-category misclassification probabilities of 0.16–0.24 and to the interview measure a context-driven threshold shift with noise standard deviation close to one.

# 3 The cohort study as setting

The data come from the Individual Development and Adaptation (IDA) longitudinal research program, which follows the school-year cohort that attended grade three in a midsize town in middle Sweden in the school year 1964–65, together with about 370 in-migrant children who joined the cohort's grade-level classes through grade nine and were integrated as full program members (Bergman, Andershed, Meehan and Andershed, 2018; Magnusson, 1988; Magnusson, Dunér and Zetterblom, 1975). Cumulative program membership is 1,392 children (710 boys, 682 girls). Most members were born in 1955, but the cohort is defined by the school year of grade three, not by birth year. Data collection covered whole school classes, so classroom instruments reached in-migrants as they joined; the grade-three instruments exist only for the original 1965 cohort.

The informant pairing also implies a context contrast. The child item was answered in a supervised school setting, while the parent questionnaire was completed at home and returned to the research program. These contexts may differ in audience, privacy, and perceived stakes; the data do not record who else was present when parents answered, or how private children felt in the classroom, so the analysis treats "context" effects as bundled with informant perspective differences.

Table 1 summarizes the design: at each of grades three, six, and nine, the child answered a classroom questionnaire item on satisfaction with school, and the parents answered a home-questionnaire item about the same child's satisfaction with school. Item wording, response scales, and per-item sample sizes are shown; the joint child–parent samples are 877 (grade three), 1,001 (grade six), and 877 (grade nine).

# 4 Design mapping and hypotheses

## 4.1 Mapping the two quasi-experiments

The parent item is, for practical purposes, a constant instrument: five fully labeled evaluative categories at all three waves, with only trivial wording drift. The child item changes format at every wave: binary at grade three, five evaluative categories at grade six, five frequency-anchored categories at grade nine; and the child's secondary "fun in school" item changes

Table 1 The informant design: parallel school-satisfaction items in the child and parent questionnaires

| Grade (year) | Child self-report (classroom) | Parent questionnaire (home) | Joint $N$ |
|---|---|---|---|
| 3 (1965) | "Are you satisfied with school?"; binary: yes / no ($N$=943) | "Is the pupil satisfied with school?"; five categories, all verbally labeled, from "perfectly great" to "not at all" ($N$=952) | 877 |
| 6 (1968) | "Do you enjoy school?"; five categories, all verbally labeled, evaluative anchors from "extremely well" to "do not thrive" ($N$=1,037) | "How does the child like school?"; five categories, all verbally labeled, same anchors as grade three ($N$=1,065) | 1,001 |
| 9 (1971) | "Have you been satisfied with school?"; five categories, all verbally labeled, frequency anchors from "nearly always" to "nearly never" ($N$=1,089) | "Has the pupil been satisfied with school?"; five categories, same evaluative anchors as earlier waves ($N$=935) | 877 |

*Note:* Cohort universe $N$=1,392. Verbatim Swedish response labels and source variable names are given in Appendix Table A1. The child's secondary "fun in school" item changed from four categories anchored "always . . . never" (grade three) to five categories anchored "nearly always . . . nearly never" (grade six).

from four to five categories with re-anchored labels between grades three and six. The IDA analog of the British 1991–92 relabeling is therefore a *within-informant* format change on the child side, with the parent series available as a control series measured on an unchanged instrument; question format is known to shape answers independently of content (Schwarz, 1999). Unlike the British case, every cross-wave contrast here confounds the format change with three years of aging and with real change in school satisfaction; the labeling analyses are correspondingly framed as descriptive. The two families of evidence divide the identification burden: the within-informant format contrasts involve one informant in one context, so perspective differences cannot generate them; the child–parent contrasts bound a mixture of context and perspective effects. They are read accordingly throughout.

The analog of the interview versus self-completion contrast is the informant–context pair itself: the same construct, answered in the same season, once by the child in a supervised classroom and once by the parents at home. The two contexts differ in audience; a parent describing the child's schooling to the school-affiliated research program faces the "put on a good show for the visitor" incentive that Conti and Pudney (2011) document for household interviews (Smith, 1979; Tourangeau and Yan, 2007), and in perspective, which is the non-separable informant component discussed throughout. The grade-six pair, where both informants face fully labeled five-point evaluative scales, is the design centerpiece: at that wave, differences in scale format are smallest and the informant–context contrast is cleanest. The two grade-six scales are similar but not identical: the child's middle category ("ganska bra," quite good) is mildly positive where the parent's ("varken eller," neither/nor) is neutral, an anchor difference the latent model of Section 6.4 probes directly.

## 4.2 Hypotheses

The article re-tests the two specific results of Conti and Pudney (2011) and their downstream implication, on new data; per the replication design there is one central hypothesis and two companions, each tied to a specific original finding and a specific exhibit.

**H1 (format and anchors).** Changes in the child item's response format and anchoring across waves are associated with changes in the shape of the child response distribution exceeding the contemporaneous change in the parent distribution (the control series), and the excess differs by sex. This parallels the original labeling result (their Figure 1 and Table 2), but is interpreted descriptively because format changes coincide with aging and because the "control" assumes comparable parental information and reporting across waves. The refuting pattern is a child-series distributional change no larger than the parent-series change. Tested in Table 2.

**H2 (informant/setting and non-classical error; central).** The two reports of the same construct diverge systematically rather than classically: the child–parent discrepancy is associated with the child's achievement, behavior, peer standing, and classroom context, differently for boys and girls; and in an integrative latent model the parent report is consistent with a systematic threshold shift (capturing response conditions and informant perspective) while the child report carries extreme-category misclassification. This parallels their Tables 3 and 5. The refuting pattern is a discrepancy unrelated to observables with misclassification and shift parameters near zero. Tested in Table 4 and Appendix Table A3.

**H3 (propagation into estimation).** Coefficients of a standard satisfaction equation differ across informant series: in particular, achievement and ability coefficients estimated from child reports differ from those estimated from parent reports, by sex. This re-tests their Table 4 contrast. The refuting pattern is coefficient equality across the two report series. Tested in Table 3.

# 5 Data and measures

## 5.1 Outcomes

All satisfaction items are re-oriented so that higher values indicate more satisfaction. Three versions are used: the native scales; a harmonized three-category version (low = original categories four–five, middle = three, high = one–two) for cross-wave work; and a harmonized binary version (satisfied = original categories one–two; the grade-three child item is already binary), which is available in every informant–wave cell and underlies the three-wave panel probit for the child series. The grade-three child item cannot support the three-category harmonization; this is a documented limitation. Verbatim labels, source variables, and recodes are in Appendix Table A1.

## 5.2 Covariates

Covariates follow the original study's logic of observable characteristics that could shift either true satisfaction or reporting behavior: sex; birth year in the cohort's three-level grouping (born 1954 or earlier, 1955, 1956 or later); cognitive ability from the grade-appropriate test battery (grade-three and grade-six total scores of the differential ability battery, grade-eight intelligence-test stanine sum), standardized; school achievement (grade-three standardized-test score in Swedish and mathematics, grade-six mark average in the same subjects, grade-nine leaving-grade average), standardized; teacher ratings of aggression, concentration, and school fatigue (grades three and six), standardized; standardized sociometric peer status; the family's social group and educational group from the grade-three parent questionnaire; a single-parent-household indicator from the 1970 census family-type code (couple versus single parent; measured at the cohort's grade eight, so two to three years after the grade-six outcome); and the classroom context (i.e., class size and the percentage of girls in the class) from the cohort's canonical class structure. With the exception of household structure (measured in the 1970 census) and the grade-eight cognitive test used as the grade-nine ability proxy, covariates are measured in the same wave as the satisfaction item. School classes (48, 46, and 47 classes at the three waves) are the sampling and clustering units.

### 5.3 Samples and missingness

Estimation samples are complete cases per model; the cohort universe is never re-based. Grade-three items exist only for the original 1965 cohort (1,031 children); the grade-six and grade-nine instruments covered the enlarged cohort. Item non-response is higher in the parent questionnaire at grade nine (935 of 1,392) than in the classroom instrument (1,089), and it is selective: the 212 children with a grade-nine self-report but no parent report describe themselves as less satisfied than those with both reports (means 3.39 versus 3.56 on the reversed five-point scale, Welch $p = 0.016$); at grade six the corresponding contrast is negligible (36 children, $p = 0.82$). Appendix Table A5 reports the contrast and a response-weighted sensitivity for the grade-nine informant comparison. Descriptive statistics by sex are in Appendix Table A2.

# 6 Results

### 6.1 Distributions and Kullback–Leibler differentials

Figure 1 plots the grade-six distributions, where the wave where both informants answered fully labeled five-point scales. Two features stand out. First, parents concentrate on the conventional positive category "good": 57 percent of boys' parents and 58 percent of girls' parents choose it, against 38 and 42 percent of the children themselves. Second, children use the interior and lower categories far more: 15 percent of boys and 9 percent of girls place themselves in the two least-satisfied categories, against 6 and 3 percent of their parents. The same concentration difference appears at grades three and nine (Appendix Figures A1 and A2). At grade three the contrast is starker still: 90 percent of boys and 97 percent of girls answer "yes" to the binary child item, while only 27 and 31 percent of parents choose the top category; the parents' fully labeled scale spreads what the child's binary format compresses.

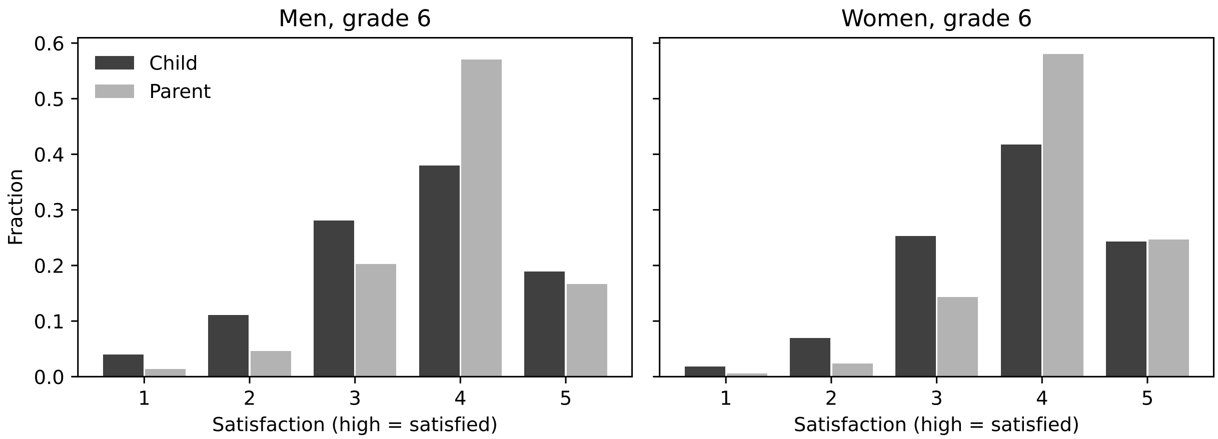


Figure 1 Child and parent response distributions, grade six, by sex

Table 2 reports Kullback–Leibler differentials (Kullback and Leibler, 1951) computed as in the original study,

$$K \;=\; \sum_j \hat{\pi}_b(j)\,\ln\!\left(\frac{\hat{\pi}_b(j)}{\hat{\pi}_a(j)}\right), \tag{1}$$

where $\hat{\pi}_b$ is the baseline distribution (the parent report within waves; the earlier wave within informants) and $\hat{\pi}_a$ the comparison distribution. Standard errors come from 1,000 bootstrap replications clustered at the school class. The support (binary, three-category, or five-category) varies across blocks, so the most direct comparisons are within a block where the support is held fixed.

Block A gives the within-wave informant contrasts. The divergence is largest at grade six: 0.090 for boys and 0.082 for girls, the wave where the two instruments are most comparable, so gross format differences (number of categories, labeled versus unlabeled) cannot account for it; the interior anchors do differ in tone (the child's middle category is mildly positive, the parent's neutral), so some divergence at the scale's center may reflect anchor semantics, a possibility the latent model probes below. The divergence is smallest for girls at grade nine (0.013), where girls' own frequency-anchored reports come closest to their parents' evaluative reports; the female–male difference at that wave ($-0.036$, standard error 0.021) has the opposite sign of the grade-three difference (+0.031, standard error 0.049). Because grade-nine parent response is selective on the child's own report (Section 5), the grade-nine contrast was re-estimated weighting the joint pairs by the inverse probability of parent response estimated on observables: the differentials are essentially unchanged (0.044 to 0.040 for boys and 0.024 to 0.025 for girls on the covariate-complete sample; Appendix Table A5). For comparison, the original study's labeling change produced differentials of 0.105 (men) and 0.069 (women) against a within-design benchmark of 0.005: the informant contrasts here are of the same order of magnitude as the British design effects, not of the benchmark.

Block B contains the H1 test. Between grades six and nine the child instrument changed anchors (evaluative to frequency) while the parent instrument stayed fixed. For boys, the child distribution moves twice as much as the parent distribution (0.080 versus 0.039; excess +0.041, standard error 0.036); for girls the pattern reverses (0.033 versus 0.096; excess −0.062, standard error 0.035). H1's directional claim is therefore not supported in a uniform way: the format change is associated with excess distributional movement for boys, while for girls the parent distribution itself shifts, consistent with parents updating on daughters' adolescent schooling, but equally consistent with age effects the design cannot remove. The control-series comparison also assumes that parents' information about the child's school life is equally current at both waves; if parental knowledge thins between ages 13 and 16, the parent series understates true change and the boys' positive excess is overstated for the same reason the girls' negative excess may reflect information rather than instrument. Block C repeats the exercise on the "fun in school" item, whose four-category "always ... never" scale became a five-category "nearly always ... nearly never" scale between grades three and six: the distributional movement (0.125 for boys, 0.096 for girls, on a common three-category support) is the largest in the table, about three times the parent-series movement over the same years (0.041 and 0.015), and, as in the original study, where the fully labeled redesign moved women more, the softened anchors moved the shares markedly out of the old absolute extreme "always."

Table 2 Kullback–Leibler differentials between response distributions

| Block | Contrast | Men | Women |
|---|---|---|---|
| A | G3 (binary) | 0.047 (0.026) | 0.078 (0.041) |
| A | G6 (native 5-pt) | 0.090 (0.024) | 0.082 (0.021) |
| A | G9 (native 5-pt) | 0.049 (0.018) | 0.013 (0.012) |
| B | Child G6->G9 | 0.080 (0.030) | 0.033 (0.021) |
| B | Parent G3->G6 | 0.041 (0.020) | 0.015 (0.011) |
| B | Parent G6->G9 | 0.039 (0.018) | 0.096 (0.028) |
| B | Excess child change G6->G9 | 0.041 (0.036) | -0.062 (0.035) |
| C | Child fun G3->G6 | 0.125 (0.036) | 0.096 (0.025) |

*Note:* Entries are Kullback–Leibler differentials (equation 1); bootstrap standard errors in parentheses, 1,000 replications clustered at the school class. Block A: baseline = parent report, same wave (grade-three contrast on the binary support; grades six and nine on the native five-point supports). Block B: baseline = earlier wave, same informant, persons observed at both waves. Block C: "fun in school" item on a common three-category support. The "excess child change" rows report the difference between the child and parent cross-wave differentials; their standard errors treat the two sexes' bootstrap draws as independent.

## 6.2 The discrepancy between informants

At grade six, on the native five-point pair, only 45 percent of pairs agree exactly, against 32 percent expected if the two reports were independent; 19 percent of children sit above their parents and 37 percent below. At grade nine, on the harmonized three-category scale, agreement is 58 percent for boys and 69 percent for girls against 49 percent under independence, and boys sit below their parents' report more often than girls (29 versus 17 percent). The independence benchmark is used only as a descriptive calibration of how much agreement exceeds what would arise mechanically from two unrelated marginals, not as a behavioral null. The original study reports 45 percent agreement on its seven-point scales; exact agreement is mechanically more likely on coarser scales, so the similarity of the two numbers is read against the independence benchmark rather than as a direct parallel (Appendix Table A5, Panel C). A multinomial logit for the three states (equal as base; child above; child below), with marginal effects at means and school-class-clustered standard errors, provides the analog of the original Table 3; marginal effects are reported at means to match the original presentation and are interpreted descriptively, with emphasis on signs and joint tests.

Appendix Tables A3 and A4 report the models for grades six and nine.

The discrepancy is not classical, and the evidence is read at two levels. Joint Wald tests of covariate blocks, i.e., an achievement-side block (ability, achievement, teacher ratings, peer status) and a context-side block (class size, class composition, household structure), reject zero loadings in five of the eight sex-by-wave tests: at grade six the achievement-side block matters for boys ($p < 0.001$) and, diffusely, for girls ($p = 0.010$), and the context-side block for girls ($p = 0.007$) but not boys ($p = 0.36$); at grade nine the pattern rotates, with the context-side block significant for boys ($p = 0.003$) and the achievement-side block for girls ($p = 0.020$). Individual marginal effects are exploratory: with roughly forty-four displayed effects per wave, exactly one cell survives a Benjamini–Hochberg adjustment at the five-percent level (girls' grade-nine achievement on the over-reporting margin, $-0.067$; marked in the tables). Read descriptively, the individual effects give the blocks their content. For boys at grade six, one standard deviation of school achievement is associated with a 6.9-percentage-point lower probability of reporting above the parent and a 7.7-point higher probability of reporting below; teacher-rated concentration shows the same one-sided pattern (+9.2 points on reporting below); and sociometric peer status moves both margins ($-4.8$ and $+5.3$ points). For girls at grade six no single achievement-side coefficient stands out, the block is jointly significant but diffuse, while classroom context is concentrated: ten additional classmates are associated with a 28.8-point higher probability of reporting above the parent. Household structure, the analog of the original study's marital-status block, enters at grade nine, where boys from single-parent households are 8.5 points more likely to report above their parents (the context-side block's joint $p = 0.003$), with same-signed but imprecise coefficients at grade six, consistent with the family-context reporting effects the original study documents for partner presence. As in the original study, where wages, hours, and partner presence moved the interview–self-completion discrepancy for one sex and not the other, the informant discrepancy loads on observables, with different loadings by sex, though the individual-cell pattern should be read as exploratory.

The class structure permits one further check: whether the informant patterns depend on a stable classroom environment. Classmate continuity is high from grade six to grade eight (a median 88 percent of a pupil's classmates retained) but low across the grade-eight-to-grade-nine transition (median 29 percent), when elective lines re-sorted pupils. No pupil retained at least 80 percent of classmates across all three transitions, an intact "same classroom all four grades" does not exist in this cohort; and 166 of the 849 pupils observed in a class at all four grades retained at least half of their classmates at every transition. The informant-discrepancy distribution does not differ significantly between these class-stable pupils and the rest at either wave, and the grade-six informant divergence is slightly smaller, not larger, in the class-stable group (Appendix Table A8). The informant effects above are therefore not an artifact of classroom instability.

## 6.3 Random-effects ordered probit models by informant

The propagation question (H3) is whether these reporting differences change the conclusions of a standard satisfaction equation. The static random-effects ordered probit

$$Y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it}, \qquad \tilde{Y}_{it} = j \;\text{ iff }\; Y_{it}^* \in [\Gamma_{j-1}, \Gamma_j), \tag{2}$$

with $u_i$ and $\varepsilon_{it}$ independent normal, is fitted separately to the parent series (five-point scale, three waves) and the child series (harmonized binary indicator, three waves; three-category scale, grades six and nine), together with a Mundlak variant, $u_i = \bar{\mathbf{x}}_i'\boldsymbol{\gamma} + \omega_i$ (Mundlak, 1978), and a dynamic variant with lagged-response indicators and Wooldridge (2005) initial conditions. With three waves the dynamic model is weakly identified and is reported as exploratory. Standard errors are clustered at the school class. Quadrature stability was checked by re-fitting the parent static model at twelve and thirty integration points against the twenty-four-point baseline; coefficients change by less than $10^{-11}$. No causal interpretation is claimed; the coefficients summarize conditional associations, and the identifying assumption for that descriptive reading is only that the random effects and disturbances are independent of the covariates conditional on the Mundlak means. Because the parent series is on a five-point scale while the child series is binary over three waves, Table 3 is complemented by a same-support comparison on the three-category outcome for grades six and nine (Appendix Table A7). Full results are in Appendix Tables A6 and A7.

Table 3 Static random-effects models by informant and sex: selected coefficients

| Covariate | Parent: men | Parent: women | Child: men | Child: women |
|---|---|---|---|---|
| School achievement (z) | 0.395*** (0.066) | 0.419*** (0.069) | 0.292*** (0.080) | 0.363*** (0.090) |
| IQ (z) | -0.076 (0.063) | -0.102 (0.064) | -0.195** (0.076) | -0.152* (0.083) |
| Peer status (z) | 0.123*** (0.044) | 0.042 (0.045) | -0.006 (0.053) | 0.078 (0.058) |
| Teacher: concentration (z) | -0.069 (0.064) | 0.005 (0.061) | -0.154* (0.080) | -0.067 (0.086) |
| Teacher: aggression (z) | -0.071 (0.053) | -0.118** (0.052) | -0.064 (0.065) | -0.196*** (0.070) |
| Class size (÷10) | 0.164 (0.128) | -0.322** (0.136) | 0.581*** (0.162) | 0.288* (0.172) |
| Percent girls (÷10) | -0.016 (0.034) | 0.009 (0.045) | -0.093** (0.038) | -0.006 (0.053) |
| Family SES group 2 | 0.060 (0.180) | -0.061 (0.183) | -0.061 (0.207) | 0.372* (0.218) |
| Family SES group 3 | -0.065 (0.182) | -0.079 (0.186) | -0.247 (0.207) | 0.327 (0.221) |
| Observations | 1,086 | 1,150 | 1,090 | 1,162 |
| Persons | 444 | 463 | 441 | 459 |

*Note:* Parent columns: random-effects ordered probit on the five-point parent report, grades three, six, and nine. Child columns: random-effects probit on the harmonized binary child report, same waves. All models include birth-year-group indicators and wave indicators. Standard errors in parentheses. *, **, ***: significance at the ten-, five-, and one-percent levels.

Three informant contrasts parallel the original study's wage result, a coefficient present in one measure and absent in the other. First, conditional on achievement, cognitive ability carries a negative coefficient only in the child's own reports ($-0.195$ for boys, $-0.152$ for girls) and is indistinguishable from zero in the parents' reports: able-but-not-high-achieving children describe themselves as less satisfied with school than either their marks or their parents suggest. Second, sociometric peer status is associated with satisfaction only in the parents' reports of boys (0.123); the boys themselves show no association; the mirror image of the original finding that social-desirability components live in the socially exposed measure. Third, the classroom variables switch informants: class size is positively associated with the child's reported satisfaction (0.581 for boys) but negatively with the parents' report for girls ($-0.322$), and the share of girls in the class is negatively associated with boys' own satisfaction ($-0.093$) but absent from every parent equation. School achievement, by contrast, is the robust common signal: 0.29–0.42 in every series and both sexes. Teacher-rated aggression is negatively associated with girls' satisfaction in both series, visible to parents and daughters alike. The Mundlak and exploratory dynamic variants (Appendix Table A6) preserve these contrasts; the lagged-response indicators are jointly significant in the parent series, indicating state persistence in parents' assessments.

## 6.4 The integrative latent model

The final exercise fits the two grade-six reports jointly as distorted indicators of one latent satisfaction variable, following the original Section IV.B. Latent satisfaction is $Y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ with standard-normal $\varepsilon$; the true category is $\tilde{Y}_i = j$ iff $Y_i^* \in [\Gamma_{j-1}, \Gamma_j)$ on shared cutpoints (both grade-six scales are fully labeled five-point evaluative scales; the anchor-semantics sensitivity below frees the parent's middle-category band, whose neutral anchor differs from the child's mildly positive one). The child report equals the true category except for misclassification of the two categories adjacent to the extremes,

$$\Pr(Y_i^C = 1 \,|\, \tilde{Y}_i = 2) = \psi_2, \qquad \Pr(Y_i^C = 5 \,|\, \tilde{Y}_i = 4) = \psi_4, \tag{3}$$

and the parent report is generated by shifted thresholds,

$$Y_i^P = k \;\text{ iff }\; Y_i^* \in [\Gamma_{k-1} - \mathbf{w}_i'\boldsymbol{\delta} - \eta_i,\; \Gamma_k - \mathbf{w}_i'\boldsymbol{\delta} - \eta_i), \qquad \eta_i \sim N(0, \sigma_\eta^2), \tag{4}$$

estimated by maximum likelihood with Gauss–Hermite quadrature. As in the original study, assigning the threshold shift to one report and the misclassification to the other is a normalization, not a substantive claim. In the present informant design this matters especially: the shift parameters $\boldsymbol{\delta}$ should be read as capturing a bundle of response conditions (audience, stakes) and informant perspective. The results are therefore interpreted at the level of fit and magnitude (whether substantial distortion is needed to reconcile the two reports), rather than as identifying which informant is "biased" in a structural sense.

Table 4 Integrative latent model of the two grade-six reports

| | Boys | Girls |
|---|---|---|
| *Distortion of the parent report ($\delta$)* | | |
| Intercept | 0.410*** (0.073) | 0.632*** (0.128) |
| School achievement (z) | 0.214 (0.152) | 0.044 (0.075) |
| IQ (z) | 0.138 (0.244) | 0.009 (0.071) |
| Family SES group 3 (low) | 0.124 (0.335) | 0.135 (0.162) |
| Standard deviation $\sigma_\eta$ | 0.957*** (0.047) | 1.001*** (0.069) |
| *Child-report misclassification probabilities ($\psi$)* | | |
| Category 2 → 1 | 0.169** (0.075) | 0.135 (0.093) |
| Category 4 → 5 | 0.221*** (0.030) | 0.339*** (0.035) |
| *Latent satisfaction equation ($\beta$)* | | |
| School achievement (z) | 0.191** (0.085) | 0.061 (0.067) |
| IQ (z) | 0.158 (0.219) | 0.018 (0.055) |
| Peer status (z) | 0.241*** (0.060) | -0.061 (0.076) |
| Family SES group 2 | 0.207 (0.260) | 0.149 (0.271) |
| Family SES group 3 | 0.109 (0.414) | 0.078 (0.117) |
| Observations | 368 | 389 |
| Log likelihood | -928.5 | -878.0 |

*Note:* Maximum-likelihood estimates of equations (3)–(4) on the grade-six child–parent pairs, by sex. Standard errors in parentheses (outer-product-of-gradients covariance; delta method for transformed parameters). *, **, ***: significance at the ten-, five-, and one-percent levels.

Table 4 shows the estimates. The child report transfers probability mass into the extreme categories with $\psi_2$ of 0.17 (boys) and 0.14 (girls) at the bottom and $\psi_4$ of 0.22 (boys) and 0.34 (girls) at the top; the same phenomenon, and almost the same range, as the original self-completion estimates of 0.157–0.236, with the largest and most precisely estimated transfer for girls at the favorable extreme (the girls' bottom-end transfer is imprecise). The parent-side threshold shift has noise standard deviation close to one (0.96 boys, 1.00 girls), matching the original $\sigma_\eta \approx 1$: joint estimation effectively down-weights the report modeled as subject to the shift/noise component. The average shift is positive (intercepts 0.41 and 0.63), implying lower implied thresholds and producing the "good"-category concentration.

The covariate structure of the shift is suggestive but imprecise: the point estimate for boys puts the shift higher for high-achieving sons (0.214, standard error 0.152), while the pooled female shift parameter is 0.222 with a standard error of 0.557. The latent-equation coefficients retain the achievement gradient for boys (0.191, standard error 0.085); the remaining latent coefficients carry large standard errors, as expected from a cross-sectional latent model with many reporting parameters, and are not emphasized.

The likelihood is multimodal. Estimates reported here come from the best of two starting configurations; the standard errors use the outer-product-of-gradients covariance because the numerical Hessian is not positive definite at the optimum. These features motivate caution in interpreting the finer covariate structure of $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$; the main emphasis is on the magnitude of the reporting parameters ($\psi_2$, $\psi_4$, and $\sigma_\eta$) and on the anchor sensitivity that follows.

The anchor-semantics sensitivity frees the parent's middle-category band: both of its boundaries shift by a common parameter $\kappa$, so the parent's neutral "neither/nor" band may occupy a different stretch of the latent scale than the child's mildly positive "quite good" band. The data place it lower, as the anchors suggest: $\kappa = -0.51$ (standard error 0.25) for boys and $-0.61$ (0.13) for girls, with a clear likelihood improvement. The reporting parameters are robust to this: the misclassification probabilities remain substantial and become similar across the sexes ($\psi_2$ of 0.20 and 0.19, $\psi_4$ of 0.25 and 0.25, for boys and girls), the noise standard deviation stays close to one (1.09 and 1.03), and the average shift attenuates (intercepts 0.26 and 0.28) as $\kappa$ absorbs the part of the parents' positivity that is the placement of their middle category. The girls' top-end transfer falls from 0.34 to 0.25: part, but only part, of the extreme-category mass reflected anchor semantics rather than response behavior.

# 7 Feasibility audit and limitations

Three departures from the original design were forced by the data, and each was resolved before estimation. First, the labeling quasi-experiment has no clean analog: the child instrument's format changes ride on three years of aging, so H1 is descriptive, disciplined by the constant parent instrument as a control series. Second, the two "modes" are two informants; the child–parent gap mixes context effects with perspective differences, and every interpretive sentence above respects that blend. Third, household structure is available only from the 1970 census, so the single-parent indicator in the discrepancy models is measured after the grade-six outcome and before the grade-nine one. Grade eight, which has a rich child questionnaire but no parent questionnaire, sits outside the informant design. The dynamic ordered probits use three waves and are weakly identified. Grade-nine parent response is positively selective on the child's own satisfaction; reweighting on observables leaves the informant contrasts unchanged, but selection on unobservables cannot be ruled out. Finally, the school-satisfaction construct itself is reported by ten-to-sixteen-year-olds; response quality of child respondents is a known function of cognitive development (Borgers, de Leeuw and Hox, 2000), and the psychological literature has long documented modest cross-informant agreement on children's adjustment (Achenbach, McConaughy and Howell, 1987; De Los Reyes and Kazdin, 2005), the contribution here is to give that discrepancy the econometric treatment of Conti and Pudney (2011) and to show it behaves like their design effects: systematic, gendered, and consequential for estimation.

# 8 Conclusion

In 1965, decades before mixed-mode designs became a standard survey concern, a Swedish cohort study administered parallel school-satisfaction items to two informants in two settings, and repeated the pairing three times over six years. Read with the informant caveat in mind, the evidence is consistent with the two broad messages of Conti and Pudney (2011) in a new construct, country, and population. Response-scale format shapes the reported distribution: the binary child item at grade three compresses what the parents' labeled scale spreads; softened frequency anchors empty the absolute extreme; and the informant divergence is largest where the scales are most comparable, making a purely mechanical scale explanation unlikely. And informant-and-setting effects are non-classical: the gap between reports is associated with achievement, teacher-rated behavior, peer status, and classroom composition, differently for boys and girls, and the choice of report changes the estimated satisfaction equation (most sharply for the ability coefficient, which is visible only in the child's own voice). The integrative model quantifies substantial misclassification and systematic threshold shifting of the same order as the original estimates. Although the cohort is historically specific, the mechanisms at issue (response-scale design and audience/perspective-driven reporting) are generic to contemporary survey measurement.

For applied work on children's subjective outcomes (a growing area in the economics of education and well-being), the practical implication mirrors the original article's: results that depend on who reported the outcome, or on how the scale was labeled, are not robust findings about children. Where informant choice is unavoidable, reporting both informants' equations, or modeling the reporting process explicitly, should be standard practice.

# References


Achenbach, Thomas M., Stephanie H. McConaughy, and Catherine T. Howell (1987) "Child/Adolescent Behavioral and Emotional Problems: Implications of Cross-Informant Correlations for Situational Specificity," *Psychological Bulletin*, 101 (2), 213–232, 10.1037/0033-2909.101.2.213.

Bergman, Lars R., Anna-Karin Andershed, Anna Meehan, and Henrik Andershed (2018) "Individual Development and Adaptation: A Life-Span Longitudinal Program Suited for Person-Oriented Research," *Journal for Person-Oriented Research*, 4 (2), 63–77, 10.17505/jpor.2018.07.

Bertrand, Marianne and Sendhil Mullainathan (2001) "Do People Mean What They Say? Implications for Subjective Survey Data," *American Economic Review*, 91 (2), 67–72, 10.1257/aer.91.2.67.

Booth, Alison L. and Jan C. van Ours (2008) "Job Satisfaction and Family Happiness: The Part-Time Work Puzzle," *Economic Journal*, 118 (526), F77–F99, 10.1111/j.1468-0297.2007.02117.x.

Borgers, Natacha, Edith de Leeuw, and Joop Hox (2000) "Children as Respondents in Survey Research: Cognitive Development and Response Quality," *Bulletin de Méthodologie Sociologique*, 66 (1), 60–75, 10.1177/075910630006600106.

Bound, John, Charles Brown, and Nancy Mathiowetz (2001) "Measurement Error in Survey Data," in Heckman, James J. and Edward Leamer eds. *Handbook of Econometrics*, 5, 3705–3843, Amsterdam: Elsevier, 10.1016/S1573-4412(01)05006-0.

Clark, Andrew E. (1997) "Job Satisfaction and Gender: Why Are Women So Happy at Work?" *Labour Economics*, 4 (4), 341–372, 10.1016/S0927-5371(97)00010-9.

Conti, Gabriella and Stephen Pudney (2011) "Survey Design and the Analysis of Satisfaction," *The Review of Economics and Statistics*, 93 (3), 1087–1093, 10.1162/REST_a_00202.

De Los Reyes, Andres and Alan E. Kazdin (2005) "Informant Discrepancies in the Assessment of Childhood Psychopathology: A Critical Review, Theoretical Framework, and Recommendations for Further Study," *Psychological Bulletin*, 131 (4), 483–509, 10.1037/0033-2909.131.4.483.

Halpern, Diane F. (2000) *Sex Differences in Cognitive Abilities*, Mahwah, NJ: Erlbaum, 3rd edition.

Kullback, Solomon and Richard A. Leibler (1951) "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22 (1), 79–86, 10.1214/aoms/1177729694.

Magnusson, David (1988) *Individual Development from an Interactional Perspective: A Longitudinal Study*, Hillsdale, NJ: Erlbaum.

Magnusson, David, Anders Dunér, and Göran Zetterblom (1975) *Adjustment: A Longitudinal Study*, Stockholm: Almqvist and Wiksell.

Mundlak, Yair (1978) "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46 (1), 69–85, 10.2307/1913646.

Schwarz, Norbert (1999) "Self-Reports: How the Questions Shape the Answers," *American Psychologist*, 54 (2), 93–105, 10.1037/0003-066X.54.2.93.

Smith, Tom W. (1979) "Happiness: Time Trends, Seasonal Variations, Intersurvey Differences, and Other Mysteries," *Social Psychology Quarterly*, 42 (1), 18–30.

Tourangeau, Roger and Ting Yan (2007) "Sensitive Questions in Surveys," *Psychological Bulletin*, 133 (5), 859–883, 10.1037/0033-2909.133.5.859.

Wooldridge, Jeffrey M. (2005) "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 20 (1), 39–54, 10.1002/jae.770.

# Appendix A. Additional tables and figures

Table A1 Source items, verbatim response labels, and recodes

| Wave | Variable (informant) | Verbatim Swedish response labels | Recode |
|---|---|---|---|
| G3 1965 | `ek337` (child) | 1 JA / 2 NEJ | yes=1, no=0 |
| G3 1965 | `ek34` (child) | 1 ALLTID / 2 OFTA / 3 NÅGON GÅNG / 4 ALDRIG | reversed, high=often fun |
| G3 1965 | `fe310` (parent) | 1 ALLDELES UTMÄRKT / 2 BRA / 3 VARKEN BRA EL DÅLIGT / 4 INTE RIKTIGT SÅ BRA / 5 INTE ALLS | reversed, high=satisfied |
| G6 1968 | `ek620` (child) | 1 UTMÄRKT BRA / 2 BRA / 3 GANSKA BRA / 4 INTE SÄRSKILT BRA / 5 TRIVS INTE BRA | reversed |
| G6 1968 | `ek63` (child) | 1 NÄSTAN ALLTID / 2 OFTA / 3 NÅGON GÅNG / 4 INTE SÅ OFTA / 5 NÄSTAN ALDRIG | reversed |
| G6 1968 | `fe614` (parent) | 1 ALLDELES UTMÄRKT / 2 BRA / 3 VARKEN ELLER / 4 INTE RIKTIGT BRA / 5 INTE ALLS | reversed |
| G9 1971 | `st923` (child) | 1 NÄSTAN ALLTID / 2 FÖR DET MESTA / 3 IBLAND / 4 SÄLLAN / 5 NÄSTAN ALDRIG | reversed |
| G9 1971 | `fe97` (parent) | 1 ALLDELES UTMÄRKT / 2 BRA / 3 VARKEN ELLER / 4 INTE SÅ BRA / 5 INTE ALLS | reversed |

*Note:* Harmonized versions: three categories (high = original 1–2, middle = 3, low = 4–5) and binary (satisfied = original 1–2; grade-three child item: yes). No sentinel codes occur in any item.

Table A2 Descriptive statistics

| | All | | | Women | Men | W−M |
|---|---|---|---|---|---|---|
| Variable | Mean | SD | $N$ | Mean | Mean | Diff. |
| Child: satisfied with school, grade 3 (0/1) | 0.94 | 0.24 | 943 | 0.97 | 0.90 | 0.07*** |
| Parent: pupil satisfied with school, grade 3 (1–5) | 4.10 | 0.73 | 952 | 4.16 | 4.04 | 0.13*** |
| Child: enjoys school, grade 6 (1–5) | 3.68 | 1.00 | 1037 | 3.79 | 3.57 | 0.22*** |
| Parent: child likes school, grade 6 (1–5) | 3.92 | 0.78 | 1065 | 4.02 | 3.82 | 0.21*** |
| Child: satisfied with school, grade 9 (1–5) | 3.53 | 0.93 | 1089 | 3.68 | 3.38 | 0.30*** |
| Parent: pupil satisfied with school, grade 9 (1–5) | 3.65 | 0.90 | 935 | 3.70 | 3.60 | 0.10 |
| Child: fun in school, grade 3 (1–4) | 3.16 | 0.81 | 964 | 3.31 | 3.02 | 0.28*** |
| Child: fun in school, grade 6 (1–5) | 3.65 | 1.06 | 1036 | 3.87 | 3.43 | 0.44*** |
| IQ, grade 3 (z) | 0.00 | 1.00 | 929 | -0.01 | 0.01 | -0.02 |
| IQ, grade 6 (z) | 0.00 | 1.00 | 1102 | 0.01 | -0.01 | 0.02 |
| IQ, grade 8 (z) | 0.00 | 1.00 | 1144 | 0.13 | -0.13 | 0.25*** |
| Standard test Swedish + mathematics, grade 3 (z) | -0.00 | 1.00 | 958 | 0.09 | -0.09 | 0.18*** |
| Grade marks Swedish + mathematics, grade 6 (z) | 0.00 | 1.00 | 1092 | 0.15 | -0.15 | 0.29*** |
| Leaving-grade mean, grade 9 (z) | 0.00 | 1.00 | 1160 | 0.21 | -0.21 | 0.42*** |
| Peer status, grade 3 (z) | -0.00 | 1.00 | 1017 | -0.00 | 0.00 | -0.00 |
| Peer status, grade 6 (z) | -0.00 | 1.00 | 1078 | -0.00 | 0.00 | -0.00 |
| Family social group, grade 3 (1–3) | 2.39 | 0.64 | 915 | 2.38 | 2.40 | -0.02 |
| Class size, grade 3 | 22.76 | 4.15 | 1031 | 22.99 | 22.52 | 0.47* |
| Class size, grade 6 | 26.66 | 4.94 | 1105 | 26.94 | 26.38 | 0.56* |
| Class size, grade 9 | 26.57 | 4.78 | 1175 | 27.30 | 25.87 | 1.43*** |

*Note:* Cohort universe $N$=1,392 (710 men, 682 women). Satisfaction items re-oriented, high = satisfied. The W−M column reports the female–male mean difference with a Welch test. *, **, ***: significance at the ten-, five-, and one-percent levels.

Table A3 Determinants of the child–parent discrepancy, grade six: marginal effects from a multinomial logit

| | Pr(child > parent) | | Pr(child < parent) | |
|---|---|---|---|---|
| Covariate | Men | Women | Men | Women |
| IQ (z) | -0.018 (0.036) | -0.053* (0.032) | -0.011 (0.043) | 0.044 (0.046) |
| Achievement (z) | -0.069** (0.029) | 0.009 (0.037) | 0.077* (0.041) | -0.032 (0.048) |
| Teacher: aggression (z) | 0.000 (0.033) | -0.015 (0.025) | -0.009 (0.038) | 0.030 (0.038) |
| Teacher: concentration (z) | -0.062 (0.041) | 0.005 (0.037) | 0.092* (0.049) | -0.015 (0.046) |
| Teacher: school fatigue (z) | 0.008 (0.035) | 0.016 (0.032) | 0.004 (0.056) | 0.043 (0.049) |
| Peer status (z) | -0.048* (0.026) | 0.026 (0.025) | 0.053* (0.030) | -0.029 (0.032) |
| Family SES group 2 (middle) | -0.051 (0.072) | 0.054 (0.091) | -0.043 (0.081) | -0.058 (0.088) |
| Family SES group 3 (low) | -0.017 (0.082) | 0.094 (0.103) | -0.053 (0.100) | -0.077 (0.111) |
| Single-parent household | 0.061 (0.059) | 0.059 (0.059) | -0.124 (0.081) | -0.115 (0.098) |
| Class size (÷10) | 0.097 (0.081) | 0.288*** (0.100) | 0.033 (0.124) | -0.166 (0.154) |
| Percent girls in class (÷10) | -0.008 (0.042) | -0.011 (0.059) | -0.054 (0.052) | -0.007 (0.065) |
| Observations | 364 | 385 | 364 | 385 |
| Share equal (base) | 0.464 | 0.444 | | |
| Share child > parent | 0.198 | 0.184 | | |
| Share child < parent | 0.338 | 0.371 | | |

*Note:* Native five-point pair; base category: child = parent. Marginal effects at means; standard errors in parentheses, clustered at the school class. Birth-year-group indicators included but not shown. *, **, ***: significance at the ten-, five-, and one-percent levels; †: survives a Benjamini–Hochberg adjustment at the five-percent level across the wave's displayed battery. Joint Wald tests that all achievement-side effects (ability, achievement, teacher ratings, peer status) are zero: $p < 0.001$ boys, $p = 0.010$ girls; context-side (class size, percent girls, single parent): $p = 0.357$ boys, $p = 0.007$ girls.

Table A4 Determinants of the child–parent discrepancy, grade nine: marginal effects from a multinomial logit

| | Pr(child > parent) | | Pr(child < parent) | |
|---|---|---|---|---|
| Covariate | Men | Women | Men | Women |
| IQ (z) | -0.023 (0.021) | 0.027 (0.018) | 0.023 (0.042) | 0.016 (0.030) |
| Achievement (z) | -0.054** (0.027) | -0.067***† (0.019) | -0.024 (0.046) | -0.038 (0.036) |
| Teacher: aggression (z) | 0.034* (0.020) | 0.016 (0.024) | -0.017 (0.040) | 0.003 (0.031) |
| Teacher: concentration (z) | -0.001 (0.025) | -0.071** (0.032) | -0.023 (0.054) | 0.003 (0.047) |
| Teacher: school fatigue (z) | -0.059** (0.029) | 0.035 (0.026) | 0.039 (0.052) | 0.031 (0.053) |
| Peer status (z) | -0.012 (0.019) | 0.005 (0.011) | 0.014 (0.029) | -0.015 (0.026) |
| Family SES group 2 (middle) | -0.030 (0.088) | -0.009 (0.063) | 0.136 (0.149) | -0.080 (0.087) |
| Family SES group 3 (low) | -0.016 (0.074) | -0.031 (0.071) | 0.213 (0.171) | -0.064 (0.092) |
| Single-parent household | 0.085** (0.038) | -0.009 (0.053) | -0.196 (0.152) | -0.007 (0.063) |
| Class size (÷10) | -0.046 (0.039) | 0.064 (0.058) | -0.110 (0.083) | 0.023 (0.048) |
| Percent girls in class (÷10) | 0.015** (0.008) | 0.003 (0.011) | 0.012 (0.020) | 0.002 (0.009) |
| Observations | 277 | 291 | 277 | 291 |
| Share equal (base) | 0.621 | 0.687 | | |
| Share child > parent | 0.123 | 0.141 | | |
| Share child < parent | 0.256 | 0.172 | | |

*Note:* Harmonized three-category pair; base category: child = parent. Marginal effects at means; standard errors in parentheses, clustered at the school class. Birth-year-group indicators included but not shown. *, **, ***: significance at the ten-, five-, and one-percent levels; †: survives a Benjamini–Hochberg adjustment at the five-percent level across the wave's displayed battery. Joint Wald tests that all achievement-side effects are zero: $p = 0.103$ boys, $p = 0.020$ girls; context-side: $p = 0.003$ boys, $p = 0.846$ girls.

Table A5 Parent-response selection, response-weighted sensitivity, and chance agreement

| | | | | | |
|---|---|---|---|---|---|
| *Panel A: parent-questionnaire response and the child's own report* | | | | | |
| Wave | $N$ resp. | Child mean | $N$ no rep. | Child mean | Welch $p$ |
| G6 | 1001 | 3.68 | 36 | 3.72 | 0.821 |
| G9 | 877 | 3.56 | 212 | 3.39 | 0.016 |
| *Panel B: grade-9 informant Kullback–Leibler differential, response-weighted* | | | | | |
| Sex | $N$ | Unweighted | IPW | (bootstrap se) | |
| Men | 292 | 0.044 | 0.040 | (0.021) | |
| Women | 308 | 0.024 | 0.025 | (0.020) | |
| *Panel C: observed vs chance exact agreement* | | | | | |
| Wave | $N$ | Observed | Under independence | | |
| G6 (native 5-pt) | 1001 | 0.448 | 0.324 | | |
| G9 (harmonized 3-cat) | 877 | 0.635 | 0.489 | | |

*Note:* Panel A compares the child's own report (reversed scale, high = satisfied) between children whose parents did and did not return the questionnaire. Panel B re-estimates the grade-nine informant Kullback–Leibler differential weighting joint pairs by the inverse probability of parent response from a logit on sex, ability, achievement, family social group, and household structure, on the covariate-complete sample; bootstrap standard errors clustered at the school class. Panel C compares observed exact agreement with the level expected if the two reports were independent.

Table A6 Random-effects models of the two report series: static, Mundlak, and dynamic variants

| | Parent report (5-point REOP) | | | Child report (binary RE probit) | | |
|---|---|---|---|---|---|---|
| Covariate | Static | Mundlak | Dynamic | Static | Mundlak | Dynamic |
| School achievement (z) | 0.408*** (0.047) | 0.426*** (0.064) | 0.569*** (0.088) | 0.325*** (0.059) | 0.434*** (0.087) | 0.560*** (0.124) |
| IQ (z) | -0.095** (0.045) | -0.044 (0.075) | -0.116 (0.098) | -0.167*** (0.056) | -0.011 (0.101) | -0.113 (0.124) |
| Peer status (z) | 0.088*** (0.031) | 0.138*** (0.044) | 0.147** (0.058) | 0.040 (0.039) | 0.164*** (0.059) | 0.160** (0.071) |
| Teacher: concentration (z) | -0.032 (0.044) | -0.009 (0.065) | -0.003 (0.122) | -0.092 (0.057) | -0.143 (0.107) | -0.281 (0.178) |
| Teacher: aggression (z) | -0.092** (0.037) | -0.040 (0.057) | -0.076 (0.107) | -0.128*** (0.047) | 0.048 (0.093) | 0.010 (0.143) |
| Class size (÷10) | -0.069 (0.089) | -0.090 (0.103) | -0.087 (0.132) | 0.387*** (0.111) | 0.259* (0.134) | 0.207 (0.169) |
| Percent girls (÷10) | -0.004 (0.024) | -0.009 (0.027) | -0.016 (0.032) | -0.055** (0.027) | -0.058* (0.033) | -0.028 (0.039) |
| Female | 0.181*** (0.069) | 0.165** (0.074) | 0.048 (0.083) | 0.292*** (0.081) | 0.269*** (0.088) | 0.164 (0.112) |
| Family SES group 2 | 0.003 (0.128) | 0.000 (0.129) | 0.154 (0.140) | 0.130 (0.148) | 0.118 (0.151) | 0.117 (0.186) |
| Family SES group 3 | -0.071 (0.130) | -0.077 (0.131) | 0.112 (0.143) | 0.012 (0.149) | -0.010 (0.152) | 0.005 (0.187) |
| Lagged response = 2 | – | – | 1.083* (0.578) | – | – | – |
| Lagged response = 3 | – | – | 0.649 (0.557) | – | – | – |
| Lagged response = 4 | – | – | 1.065* (0.575) | – | – | – |
| Lagged response = 5 | – | – | 1.347** (0.619) | – | – | – |
| Lagged response (binary) | – | – | – | – | – | -0.030 (0.353) |
| Observations | 2,236 | 2,236 | 1,345 | 2,258 | 2,258 | 1,306 |
| Persons | 907 | 907 | 780 | 903 | 903 | 710 |

*Note:* Parent series: five-point random-effects ordered probit, grades three, six, and nine. Child series: random-effects probit on the harmonized binary indicator, same waves. Mundlak variants add person means of the time-varying covariates; dynamic variants add lagged-response and initial-condition indicators (grades six and nine only) and are exploratory with three waves. All models include sex, birth-year-group, family social group, and wave indicators. Person-mean (Mundlak) coefficients are estimated but not shown. Standard errors in parentheses. *, **, ***: significance at the ten-, five-, and one-percent levels.

Table A7 Three-category child series and same-sample parent series, grades six and nine

| Covariate<br>Grades 6 and 9 only | Child 3-cat static | Child 3-cat Mundlak | Parent 5-pt static |
|---|---|---|---|
| School achievement (z) | 0.360*** (0.060) | 0.392*** (0.096) | 0.544*** (0.062) |
| IQ (z) | -0.135** (0.057) | -0.054 (0.121) | -0.148** (0.058) |
| Peer status (z) | 0.034 (0.041) | 0.173** (0.070) | 0.071* (0.041) |
| Teacher: concentration (z) | -0.099 (0.063) | -0.114* (0.067) | 0.042 (0.065) |
| Teacher: aggression (z) | -0.128** (0.052) | -0.136*** (0.053) | -0.145*** (0.054) |
| Class size (÷10) | 0.299*** (0.110) | 0.091 (0.148) | -0.011 (0.116) |
| Percent girls (÷10) | -0.014 (0.025) | -0.019 (0.033) | 0.002 (0.028) |
| Female | 0.175** (0.086) | 0.145 (0.092) | 0.152* (0.088) |
| Family SES group 2 | 0.138 (0.157) | 0.117 (0.159) | 0.142 (0.161) |
| Family SES group 3 | 0.045 (0.158) | 0.031 (0.161) | 0.098 (0.164) |
| Observations | 1,465 | 1,465 | 1,404 |
| Persons | 812 | 812 | 822 |

*Note:* Random-effects ordered probits on grades six and nine. All models include sex, birth-year-group, family social group, and wave indicators. Standard errors in parentheses. *, **, ***: significance at the ten-, five-, and one-percent levels.

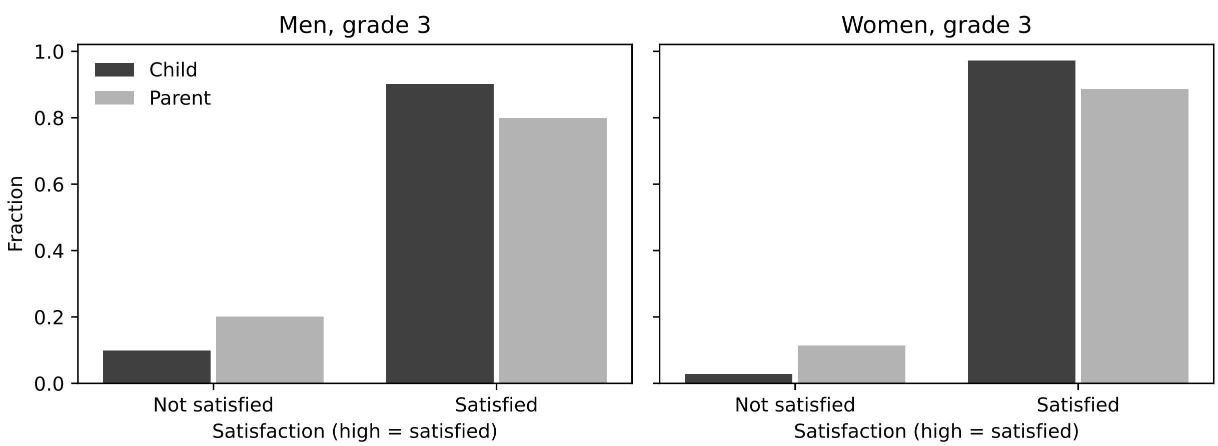


Figure A1 Child and parent response distributions, grade three, by sex

Table A8 Classmate continuity across grades and the informant discrepancy

| *Panel A: classmate retention across transitions* | | | | | |
|---|---|---|---|---|---|
| Transition | $N$ | Mean | Median | Share $\geq$0.5 | Share $\geq$0.8 |
| Grade 3 → 6 | 927 | 0.54 | 0.64 | 0.61 | 0.32 |
| Grade 6 → 8 | 1032 | 0.84 | 0.88 | 0.94 | 0.83 |
| Grade 8 → 9 | 1140 | 0.38 | 0.29 | 0.36 | 0.05 |
| Observed at all four grades | 849 | of 1,392 | | | |
| Class-stable ($\geq$0.5 every transition) | 166 | | | | |
| Class-stable ($\geq$0.8 every transition) | 0 | | | | |
| *Panel B: informant discrepancy by class stability* | | | | | |
| Wave, group | $N$ | Equal | Child > parent | Child < parent | |
| Grade 6, class-stable | 157 | 0.414 | 0.229 | 0.357 | |
| Grade 6, class-changing | 617 | 0.451 | 0.185 | 0.365 | |
| Grade 6: equality test | $p = 0.432$ | | | | |
| Grade 9, class-stable | 134 | 0.604 | 0.187 | 0.209 | |
| Grade 9, class-changing | 515 | 0.656 | 0.124 | 0.219 | |
| Grade 9: equality test | $p = 0.173$ | | | | |

*Note:* Retention is the share of a pupil's cohort classmates at the earlier grade who are again in the pupil's class at the later grade (whole-class sampling makes cohort members approximately the full class). Class-stable pupils have a class identifier at all four grades and retention of at least 0.5 at every transition. Panel B reports the child–parent discrepancy shares by stability; the equality tests are chi-squared tests of the three-state distribution across the two groups.

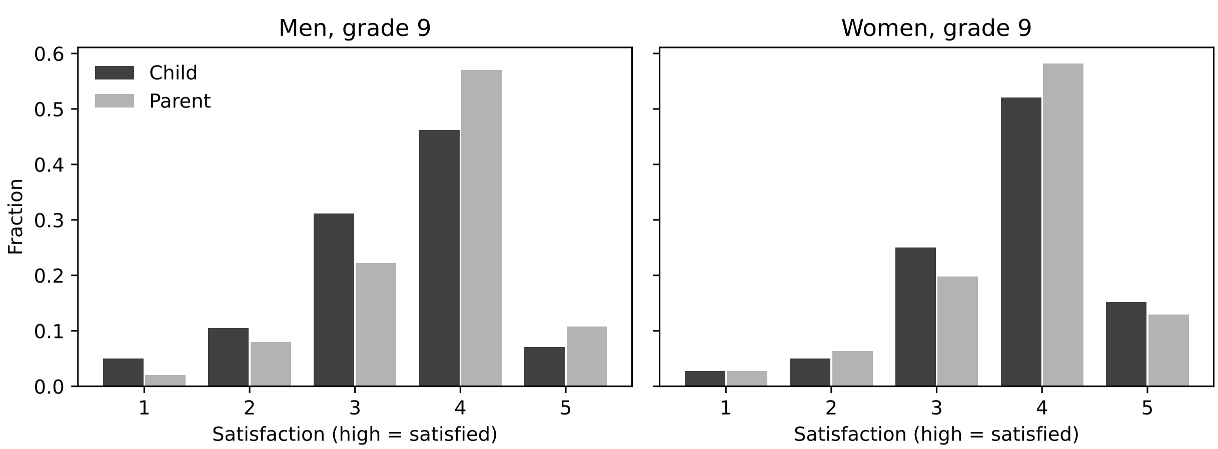


Figure A2 Child and parent response distributions, grade nine, by sex

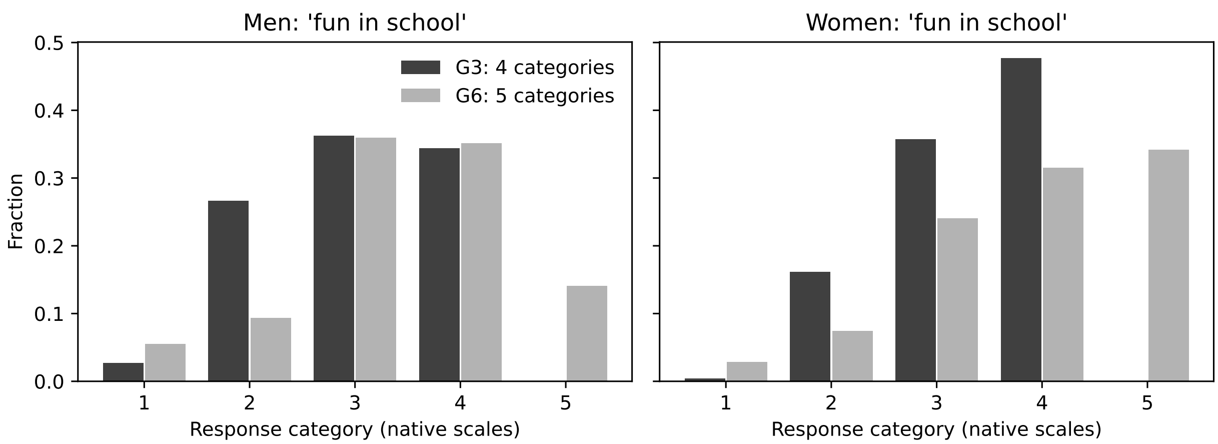


Figure A3 The “fun in school” item before and after the format change, by sex