Time-of-Survey Dependence of Apparent Educational Participation— A Demographic Perspective

DRAFT - NOT FOR CITATION OR CIRCULATION

Bilal Barakat

Wittgenstein Centre for Demography and Global Human Capital (IIASA, AW/VID, WU)
Vienna Institute of Demography (VID), Wohllebengasse 12–14, A-1040 Vienna, Austria
bilal.barakat@oeaw.ac.at

Abstract

While seemingly straightforward, many common indicators of educational participation, such as net enrolment ratios, are subject to distortions that are transparent from within a demographic framework, but generally neglected in the field of educational development. With a focus on the fact that the true school-age population is misidentified when age is measured in whole years in household surveys conducted some time after the beginning of the school year, and on the interactions of this effect with the cut-off date, population growth, and drop-out, it can be shown both analytically and through simulations that the errors induced in customary participation indicators — and potentially also in statistical analyses of the determinants of education — can be considerable. At the same time, approximate corrections are possible even with the data currently available, that is, when age is measured in whole years only. In addition, the demographic perspective adopted suggests an alternative interpretation of the out-of-school rate in terms of person-time rather than headcount.

Keywords: education statistics, indicators, household surveys, Lexis diagram, age-period-cohort

1. Introduction

A child's age plays a prominent role in the schooling process and in education statistics. It frequently determines at what time it may enter school, and at what time it must, whether it is considered under or over the 'norm' age for a given grade, and both individual age and the nominal school age range enter the calculation of enrolment rates.

While in reality individual age is a continuous, real-valued variable (and, in principle, so is the entry age, as different cut-off dates for determining it demonstrate), for most purposes in educational administration and monitoring it is recorded as an integer-valued variable. In other words, what is recorded as 'age' is frequently 'age at last birthday', or equivalently: 'age in whole years'. Crucially, this is the case both for educational administrative data forming the basis for many international collections of statistics, for reasons of succinctness, as well as for many sets of micro-data from household surveys, where, even if exact birthdays are recorded, only integer age is published for reasons of confidentiality.

For education indicators calculated from survey data, this raises the issue that some of those aged 6 *at time of survey* and not enrolled will have been only 5 years old at the beginning of school year, and are therefore not actually of school age and therefore not actually 'out of school', if the official entry primary school age range were 6 to 11 years, for example. Similarly, someone aged 12 at time of survey may have been 11 at the beginning of school year (and 6 at the beginning of his or her first school year), and is not, in fact, 'over-age'. This is an entirely different issue from that addressed by the 'adjusted NER', which deals with students observed to be *in* the relevant age range who are already enrolled in the next phase. It is also not addressed by using the gross enrolment ratio instead; while this does include the 12-year-old 'on time' children in primary, it also includes students who really are, and are known to be, outside the appropriate age range.

The present discussion focuses on the shifts observable in household surveys (including censuses), where the age of the student has a different reference date (namely the time of the survey) than the nominal age range (typically, the beginning of the school year). A related, but different phenomenon affects administrative data indicators, if the 'age' recorded there is the student's age on December 31, i.e. if calendar birth year cohorts are considered instead of school entry cohorts, or if the population figures for the denominator in the calculation of rates has a different reference date. While this also deserves attention, the household survey issue is more relevant to a general audience, who may well be in the position of wishing to calculate participation indicators for specific sub-populations (defined by socio-economic status, for instance) based on survey data, but who, in general, will not be in a position to use administrative micro-data directly.

The fact that some students will have celebrated their birthdays in the intervening time between the reference date and the collection of survey data is, of course, not a novel observation. However, to date it has not been systematically analysed, much less systematically corrected for. One of the earlier discussions can be found in a document by the UNESCO Division of Statistics (1997), where a rudimentary attempt to correct for it is made, but curiously justified purely in terms of supposedly 'late entry' instead of the natural age pro-

gression with respect to different reference dates.[1] In UNESCO's Institute for Statistics report on 'Children Out of School' (UIS, 2005), the effect is account for in a rather ad hoc and coarse way, by shifting the age range by a whole year if the survey being used was 'late in the school year' (p. 64–5). In a separate analysis of discrepancies between administrative and survey data on participation, the existence of the effect is noted, but the possible magnitude of error is assumed *a priori* to be relatively small (UIS, 2010, 43)—a premature assumption, as the following analysis will show. The World Bank, too, in its guide for it's new *ADePT Edu* online data platform, notes the existence of the effect , but no concrete solution is proposed beyond the generic exhortation that '[t]he timing and duration of household surveys relative to the school year should be taken into consideration when interpreting education indicators derived from household surveys' (Porta et al., 2011, 28), which offers no advancement over the same formulation employed seven years earlier by **?**.

The present analysis demonstrates that taking this effect and its interactions with the exact entry age, drop-out and population growth into consideration successfully is a subtle business that benefits greatly from a more systematic treatment. What is absent from the literature and documented practice are a systematic analysis of the potential magnitude of the error, of the interaction with other errors induced by population growth, drop-out, and the related, but distinct, issue of fractional *entry threshold* ages (rather than of integer *respondent* ages), and of systematic strategies for adjusting indicators to correct for these effects. The present study attempts to begin to fill this gap.

The structure of the presentation is as follows. First, a conceptual framework is presented, borrowed from demographic analysis, that facilitates the disentangling of the effects of entry age thresholds, cut-off dates, age progression and grade progression on apparent net enrolment rates. A number of interrelated but distinct distortion effects are analysed that affect the apparent net enrolment rate when these are based on integer age at the time of survey and on integer ranges for the nominal school age. Next, a general method is suggested for bounding the effect of the dominant age progression effect on standard indicators or statistical estimates. Since for the most sophisticated approach to correcting for the age progression effect only general principles can be formulated whose application depends on the specifics of the available data, it is best illustrated with a concrete country case study. Here, this is done for Indonesia, and the case study at the same time demonstrates that both the simple bounding estimates and the more sophisticated approach yield insights of policy relevance. Finally, against the backdrop of an understanding of participation as a process happening on an age-time surface, it is argued that non-participation is best understood not as a count of out-of-school children, as it commonly is, but as an attempted measurement of person-time spent in the non-enrolled state, in other words, as out-of-school child*hood*. The concluding section summarises the implications of the analysis for educational statistics and survey design.

[1] While noting elsewhere that the reference date for exceeding the school entry age in Chile and Germany (treated as exceptions) differs from the beginning of the school year, the adjustment is subsequently justified by the claim that 'since only a part of the population age 6 years enter at age 6 and the remaining population aged 6 enter when aged 7 years' (p. 25), with no distinction being made between 'aged 6 at time t_0' and 'aged 6 at time t_1'.

2. Demographic distortion effects

2.1. Enrolment on the Lexis diagram

An analysis of life events and the different ways of aggregating them by time, age, or cohort is best based on the so-called Lexis diagram. In Figure 2.1, this is adapted to the schooling context. The y-axis shows age, the x-axis time. Accordingly, individuals move along diagonal trajectories as they age over time. Parallelogram (a) represents the lifetime spent in school year 2 by those individuals who turn 9 years old, in other words, whose diagonal life path crosses the age=9 line, during that school year. Some of the time these children spend in school year 2 is at age 8, and some of it at age 9. Note that these individuals constitute a single school entry cohort, as they had the same age in completed years at the beginning of their first school year. Parallelogram (b) belongs to the same school entry cohort, but highlights the lifetime spent by its members at the age of 10, more precisely: between their 10th and 11th birthdays. Some of this time is spent in one school year, and some in the following school year, and similarly with respect to different grades. Finally, square (c) represents the lifetime of 10-year-olds in school year 2. Over the course of that school year, some children enter this area as they turn 10, while others leave it when they turn 11. As a result, two different entry cohorts contribute to (c). Without loss of generality, assume a 5-year cycle and entry of those aged 6 years at the beginning of the school year. The 'nominal age range' corresponding to this school cycle is 6–10 in whole years of age, and this is the age range that would be published in the UIS database, for example. This is true for a 'standard' student, who is *exactly* 6 years old on the first day of school and all throughout the first grade, 7 years old throughout grade 2 etc. In the Figure, the life trajectory of such a student is labeled 'Jane'. Naturally, most children turn 6 at some intermediate point during the school year preceding their entry. As a result, they turn 7 at some point during grade 1, and also turn 11 at some point during their final school year in grade 5. In the Figure, 'John' has such a trajectory. This is not because he is over-age; John could was not eligible to enter school the year before he did, because he was still only 5 at that time. The lightly shaded area marks the lifetime corresponding to the actual (normal) life spent in primary school, and, because of the above, contains some lifetime at age 11.

2.2. Age progression

The natural progression of age will lower the apparent net participation based on integer years of age at the time of survey, and do so more, the later in the school year the survey is conducted. In Figure 2.1, we can see that using the integer age school-age window, a cross-sectional survey (i.e. a vertical line) will capture some of a triangle of type (e) and miss some of a triangle of type (d). Only one triangle of each type is labelled in the diagram for sake of clarity, but each year actually contains both. Triangles of type (d) is the person-time spent by students in the final grade after their age has increased in the current school year. As a result, these individuals fall outside the nominal integer-year age range, even if they entered school at the norm age. Triangles of type (e) contain the person-time spent by children at the nominal entry age, but having only reached that age since the beginning of

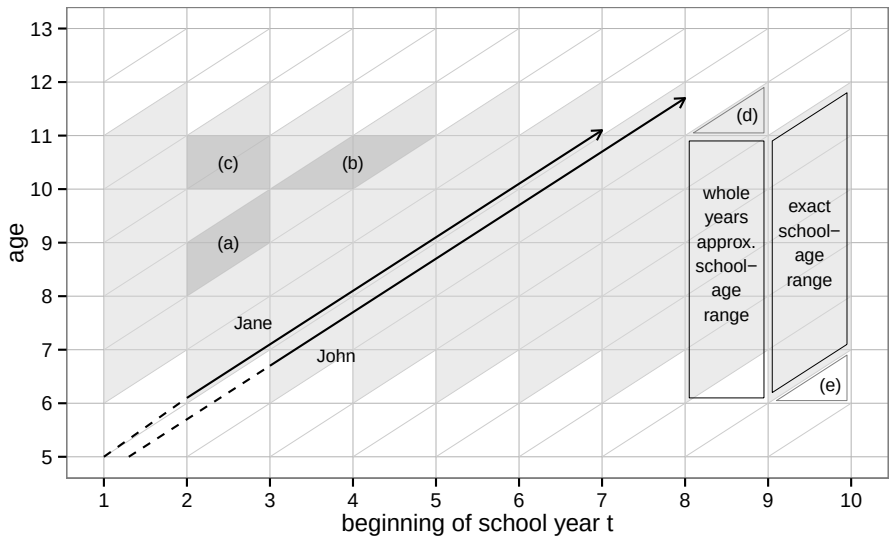

Figure 1: An educational Lexis diagramme

the school year. As a result, these individuals fall inside the nominal integer-year age range, even though there were not supposed to enter school in the current school year.

Geometrically, assuming a uniform distribution of births along the x-axis, it is intuitively clear that the effect increases linearly over the course of the school year, and that by the end of the school year, we are excluding a whole year-group of potential students from, and including a whole year-group of non-students in, the nominal school-age bracket based on integer years of age. This is clearly reflected in the results of simulating apparent net enrolment based on integer age at different times during the school year, shown in Figure 2.2, which is based on an assumption of universal entry and zero drop-out in a five-year primary cycle.

The treatment of the break between two school years deserves special comment. While the end of one school year and the beginning of the next may in general be considered to occur in the middle or at end of the long break, for now the new entry cohorts' enrolment is assumed to begin at the start of the break. The graduating class will cease being enrolled at the end of the school year proper. The actual behaviour of empirical participation indicators during the break between school years is unpredictable, but can be bounded above and below. While it is fairly clear that the completers of the final grade are no longer counted as enrolled (in that school phase), the enrolment status of the prospective new entrants is unclear. Instructions may differ between countries (compare, for example, the instructions for the Indonesia 2010 and Jamaica 2001 censuses on this point (IPUMS, 2011); they are certainly unlikely to encourage coding consistent with each other). More often than not, survey instructions do not specify the status of prospective entrants. Even when there are instructions, it is clear that the question poses much potential to be interpreted differently by different respondents. Nevertheless, it is clear that bounds can be established for present purposes by assuming either that none or all of the prospective entrants are counted as

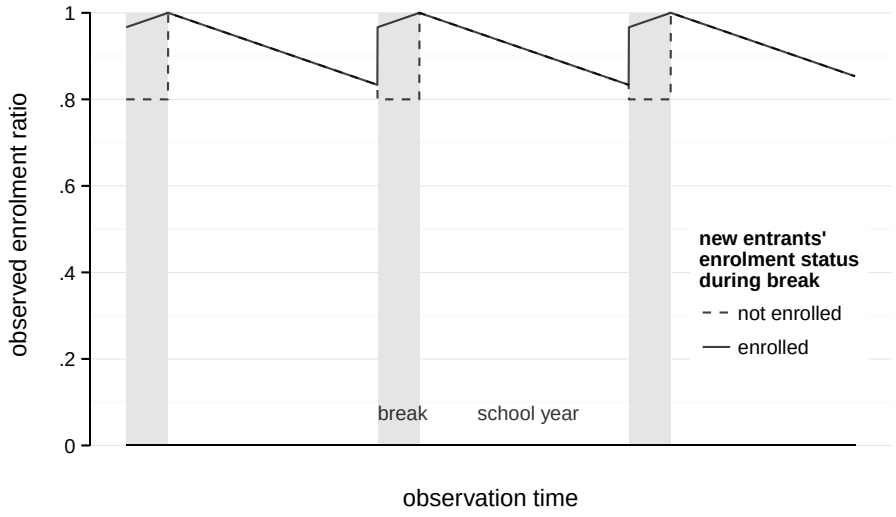

Figure 2: Apparent participation based on integer years of age

enrolled during the school break preceding their entry.

Note that the effect is not symmetrical: if they are not counted, the apparent participation rate remains at a constant reduced level during the entire break, while if they are counted, it increases monotonically to peak at the beginning of the new school year. This is because some of those who will have reached the age of 6 by the time the new school year starts, and will therefore enter, only celebrate their birthday during the break, and some of the graduates only reach age 11 during the break. As a result, over the course of the break, no-longer-enrolled graduates in the calculation of the whole-years-based participation indicator are continuously being replaced by already-enrolled new entrants.

By taking the new entry cohort to begin their enrolment at the time that the graduating cohort finishes, the 'true' enrolment rate, if entry is universal and drop-out zero, is a constant 100%. This makes for a simple baseline to which to compare the distorted estimates. If new entrants are only considered to be enrolled once the new school year has started, the 'true' enrolment rate undergoes a dip during the break, when the graduates have left, but the new entrants have not yet started, so the number of enrolled cohorts is one less than the number of grades.

Another reason for applying the convention of treating prospective entrants as enrolled (once they've crossed the age threshold) for illustrative purposes is that in 2.3 below, it makes it easier to understand how the apparent enrolment at the beginning of the school year arises geometrically than if the break were left blank, for instance.

2.3. Exact entry age threshold

A related effect, but distinct from age progression, is the shift in apparent enrolment rates based on integer age at the time of survey that occurs when the entry age threshold is actually not integer, or equivalently, if the reference date for the integer age threshold does

not coincide with the start of the school year. This is not uncommon, and there are cases both of reference dates before and after (often December 31) the beginning of the school year. Indeed, in Germany cut-off dates vary between states from June 30 to December 31, and the equivalent exact entry ages from 6 years 2 months to 5 years 8 months (bildungsserver.de, 2012) within a single country.

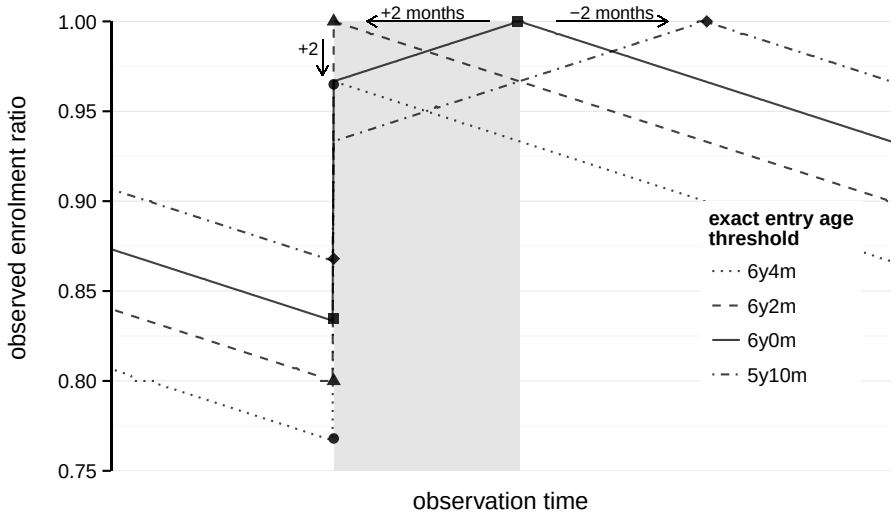

Figure 3: Exact school entry age thresholds and their effect on apparent integer-age participation

The effect on apparent participation using integer-age indicators is not simply a lateral shift of the integer age threshold pattern. As a matter of fact, if the exact entry age threshold exceeds the assumed integer entry age by more than the duration of the break between school years, an overall *downward* shift occurs, in addition to a shift in the timing of peak apparent enrolment. In other words: if the age threshold is 6.x years exactly, where .x is a greater fraction of the year than the school break occupies, then even if entry is universal and drop-out and repetition non-existent, i.e. when the true enrolment rate is 100% throughout, the apparent integer-age enrolment ratio will never reach 100%, at no point during the school year! The interaction of the exact entry age threshold with the age progression effect on apparent enrolment based on integer age at time of survey is shown in Figure 2.3.

The explanation of the effect for deviations from the integer threshold age of up to the duration of the break are straightforward, by considering at what point in time the no-longer-enrolled graduates are old enough to drop out of the integer school age range and the new, enrolled, entrants enter it. To understand the downward shift when the reference date is earlier than the beginning of the break, suppose that the entry age threshold is 6 years, with a reference date 2 months before the end of the previous school year, and 4 months before the beginning of the next. Note that for the children who reach 6 years of age between said reference date and the beginning of the break, their lifetime spent in the school age range and at the same time in an enrolled state will be strictly shorter than the duration of schooling. That is because they spend some of their time at age 6 before they

could be attending, and some of their penultimate (!) school year outside the assumed age range.

Note that the difference between an exact entry age of 6 at a reference date of June 30 for a school start on September 1, versus at a reference date of December 31, will in the latter part of the school year induce a 10 percentage point difference in the apparent enrolment rate. In the latter case of an end-of-year reference date, the apparent enrolment when the true enrolment rate is 100% drops as low as 77%.

2.4. Population growth

Unlike indicators based on administrative data, for survey-based indicators the population in the denominator comes from the same data source. The problem that population growth makes a mid-year or end-of-year population slightly out-of-sync with administrative enrolment figures does not occur. However, population growth, or, to be more precise: cohort-on-cohort growth (regardless of how the overall adult population changes), not only affects the apparent participation rate directly ([ANONYMISED]), but also affects the difference between the exact age and integer age indicators. This is shown in Figure 2.4.

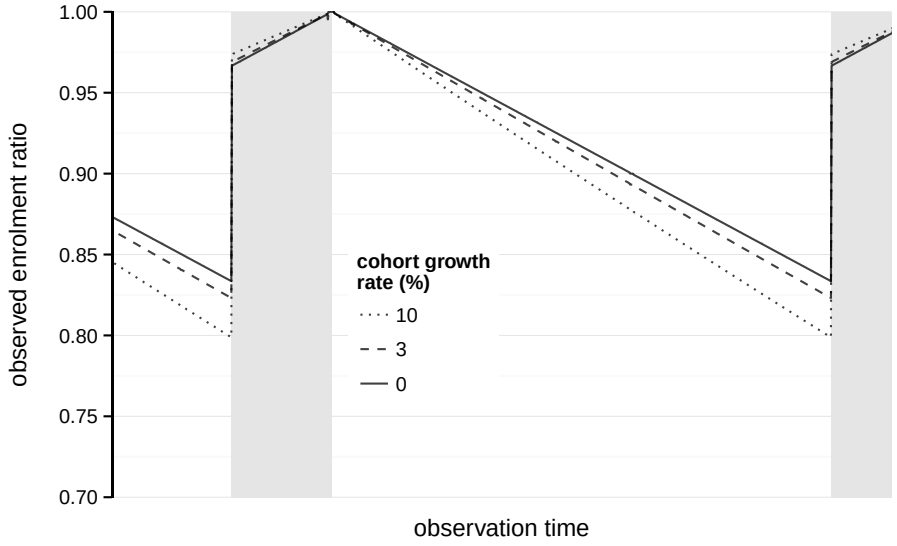

Figure 4: Cohort-on-cohort growth and its effect on apparent participation

The effect occurs because, as the school year progresses, a cross-section observing integer ages at time of survey not only gradually replaces members of a triangle of type (d) who are of actual school age by ones of type (e) who are not, but seem to be, but in addition the former, who may or may not be enrolled, and in the above simulations are enrolled, are smaller in number than the latter, who are not (unless they enrolled early, which is not assumed in the simulation). A cohort-on-cohort growth rate of 10% annually is extreme, of course, and merely serves to make the effect more visible at the natural scale.

In general, a longer school cycle, i.e. a greater number of grades and years of school-age, will compound this effect, because the birth cohorts being incorrectly excluded/included differ more in size if they are more distant in birth year.

A subtlety here is that the bias induced by population growth on the apparent enrolment rate is of opposite sign on either side of the time of peak apparent enrolment, and grows proportionally to the distance from it. Together with the effect of the exact entry age threshold discussed in Section 2.3 above, this means that an exact entry age threshold that is *above* its rounded integer value tends to *decrease* the maximal bias introduced by cohort growth (by shifting the peak closer to the middle of the school year), while an exact entry age threshold this is *below* its rounded integer value tends to *increase* the maximal growth bias.

Either way, the distortion induced by cohort growth is an order of magnitude smaller than the age progression effect, and will be ignored in the following.

2.5. The interaction with drop-out

For illustrative purposes, we distinguish drop-out over the course of the school year from drop-out between two school years, with the former presenting as a continuous decline of enrolment and the latter as stepped decline (here modelled to occur in the middle of the break). The 'peaks' in enrolment in the first half of the break appear because, by assumption, the new entrants are already counted but the drop-outs have not yet dropped out. Extreme drop-out rates shown here that serve merely to illustrate the principle of the off-setting effect. The necessity of making the distinction between drop-out during and between school years arises from the fact that the apparent enrolment due to age progression *also* presents as a continuous decline. As a consequence, in Figure 2.5, in panels A and D, which show low and high levels of continuos drop-out respectively, it may seem as if high drop-out, in addition to lowering enrolment overall, does not affect the apparent enrolment based on integer age, but does lower the exact age enrolment to match it. However, reading the panels in the sequence A, B, C, D makes the correct interpretation more obvious. Specifically, high drop-out as such decreases the estimation error in apparent net enrolment based on integer age at time of survey, because enrolment in the final grade is then anyhow minimal, so the triangle (d) which is incorrectly excluded contains mostly non-enrolled children just like the incorrectly included triangle (e) does. In addition to this effect, the exact-age and integer-age indicators *both* track continuous drop-out over the course of the year.

That these two effect off-set each other in the example is not a coincidence. As a matter of fact, it can be shown that in general the diminishment of the age-progression decline on the one hand, and the drop-out-induced decline itself on the other, will approximately cancel each other out in the integer age indicator. To see this, note on the one hand that the group of students contributing to the age-progression decline are those in the final year. Let annual cumulated drop-out be $100r\%$, and let $s = 1 - r$, and there be n grades. Then, only s^n of the oldest entry cohort are left at the end of the final year, so counting them (exact age) or not (whole years age) makes a difference, in absolute share, of:

$$\frac{1 - s^n}{n}.$$

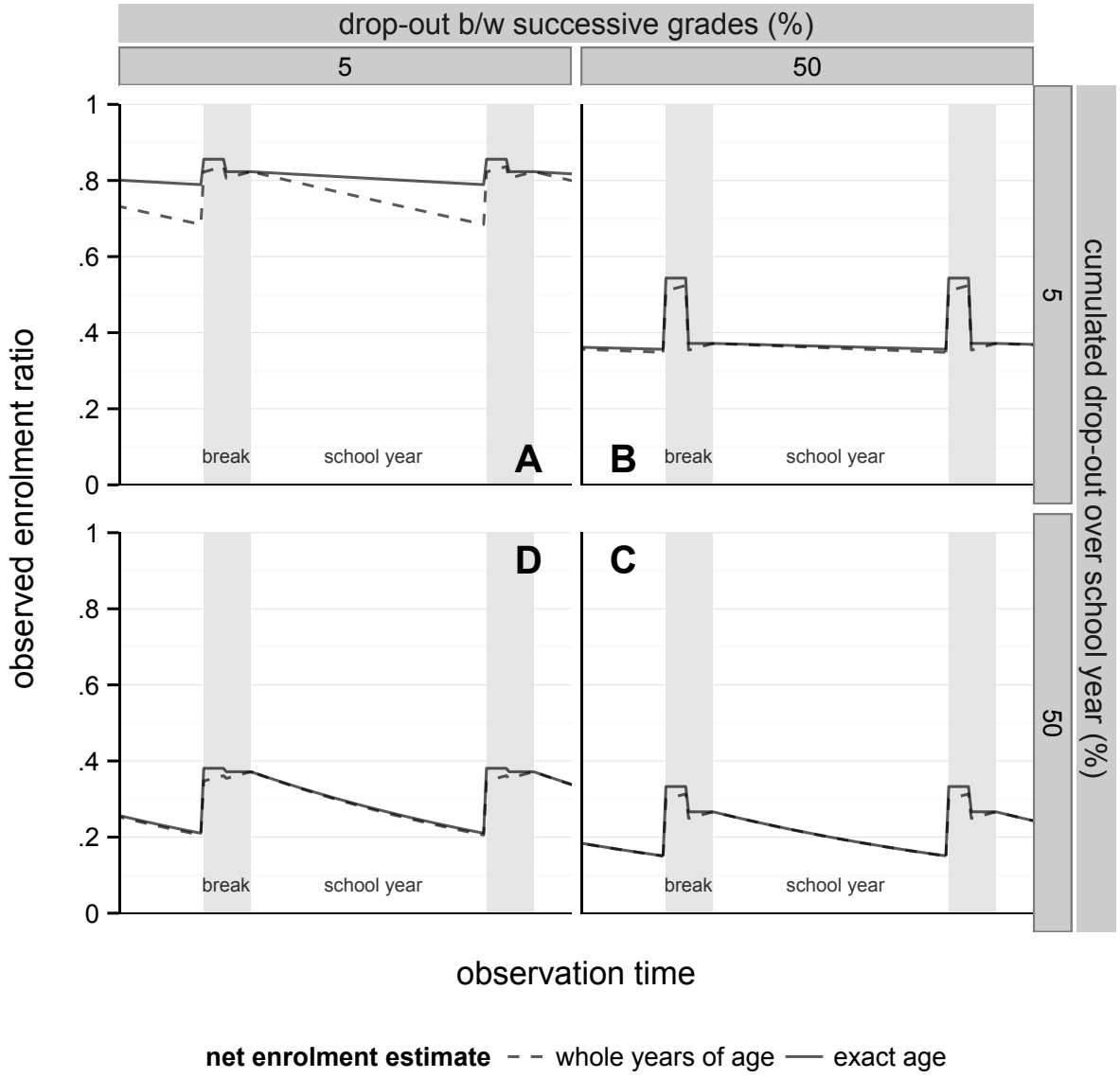

Figure 5: Interaction of drop-out and age progression

On the other hand, at the beginning of the school year, the total student population is

$$\frac{1}{n}\sum_{i=0}^{n-1} s^i,$$

and a fraction $r = 1 - s$ of these will drop out by the end of the year. So in absolute terms, the total drop-out in all grades together will likewise be

$$\frac{1}{n}(1-s)\sum_{i=0}^{n-1} s^i = \frac{1}{n}(\sum_{0}^{n-1} s^i - \sum_{1}^{n} s^i) = \frac{1-s^n}{n}.$$

In reality, of course, drop-out may not occur evenly distributed over the school year, or across grades.

3. Correcting for age progression

For an overall participation measure, the simplest solution to all the above issues, with the exception of drop-out, is simple: calculate enrolment rates—using age in whole years—restricted to the age range corresponding to the nominal ages in grades 2 through to $n - 1$, where n is the duration of the school phase. In other words, drop the first and last grades. While this does lose information, it does so only compared to the true measure, not necessarily compared to the customary measure that attempts to approximate the full age range, but is still only an approximation. In any case, such robust participation measures would still be sensitive to trends over time and to overall system performance.

However, more informative corrections for the age progression effect can be performed. For the age progression effect, this can be done with relatively little data, in contrast to drop-out, where an assumption of uniformly spaced drop-out appears highly implausible even as an approximation. For this reason, in the following, no attempt is made to correct for drop-out that occurred during the current school year (the cumulated effect of past drop-out anyhow does not need to be taken into account explicitly in performing the adjustment, it will simply automatically affect the difference the adjustment makes, by reducing the enrolment in the final grade). As the above analyses show, however, since the age progression effect is much greater than either the effect of cohort growth or of drop-out in a single year, correcting only for the former serves as a reasonable first-order approximation.

The fact that the data report age in integer years does not imply that only integer-year age ranges for dichotomous inclusion/exclusion can be considered. This latter approach appears to have been followed by UNESCO (2005), where in countries for which the underlying survey had bee conducted late in the school year, participation indicators were calculated based on a shifted age range of 7–11, say, when the nominal primary school age range was 6–10.

If we are interested in indicators for the primary cycle as a whole, say, rather than age or grade specific indicators, only the lowest and highest age group in the range need to be considered. Suppose these are 6 years for the entry age, and that we have a 6-grade cycle. The general approach is to assume or estimate the probability that a 6-year-old at time of

survey was in fact only 5 years old at the beginning of the school year and is therefore not in the exact school age range, *conditional on* his or her observed enrolment status, and similarly whether a 12-year-old at the time of survey was in fact 11 years old at the beginning of the school year, and therefore actually *is* still in the exact school age range. This probability should, again, be conditional on observed enrolment status.

The following approach is suggested. This is essentially a generalisation and elaboration on the approach that was applied in an ad hoc fashion to only a subset of countries (and justified differently) in the older report by UNESCO Division of Statistics (1997, 24). Suppose we knew the probabilities

$$p(a_{si} = a_t i|e_i) = 1 - p(a_{si} = a_{ti} - 1|e_i),^2$$

where a_s is the integer age at the beginning of the school year (unknown) and a_t is the integer age observed at survey time t, and e is an indicator of enrolment ($e = 1$) or non-enrolment ($e = 0$). The index i across individuals is dropped in the following, assuming homogeneous probabilities in the population. Then, given the numbers enrolled n_a^1 and not enrolled n_a^0 (not exponentiation!) by single years of age a, proceed as follows.

In the calculation of the primary net enrolment rate, replace terms n_6^1 in the enumerator and $n_6^1 + n_6^0$ in the denominator by $p(a_s = a_t|1)n_6^1$ and $p(a_s = a_t|1)n_6^1 + p(a_s = a_t|0)n_6^0$ respectively (to exclude the triangle of type (e)), and add $(1 - p(a_s = a_t - 1|1))n_{12}^1$ and $p(a_s = a_t - 1|1))n_{12}^1 + p(a_s = a_t - 1|0))n_{12}^0$ to the enumerator and denominator respectively (to include the triangle of type (d)).

Now note that the *unconditional* probabilities $p(a_s = a_t)$ and $p(a_s = a_t - 1)$ can be estimated purely based on the time of survey information and a uniform birth month assumption (or even better, using birth month distribution 6 years previously from census or registration data).

The impact of different assumptions about the conditional probabilities can now be compared. By Bayes' theorem in odds form, we have

$$\frac{p(a_s = a_t|1)}{p(a_s = a_t - 1|1)} = \frac{p(1|a_s = a_t)}{p(1|a_s = a_t - 1)} \times \frac{p(a_s = a_t)}{p(a_s = a_t - 1)}.$$

In real settings, we would normally expect underage children to be less likely to enrol than entry-age children, $p(1|a_s = a_t) \geq p(1|a_s = a_t - 1)$. We can therefore estimate an upper bound on the out-of-school rate by assuming that the true age is independent of enrolment status and setting $p(a_s = a_t|1) = p(a_s = a_t) = p(a_s = a_t|0)$.

Conversely, a lower bound can be calculated by assuming that enrolment occurs in perfect order according to true age. Specifically, at the entry age, we draw those children assumed to already having been of entry age at the beginning of the school year from the enrolled children as far as possible, and conversely draw the children assumed to having been below entry age at the beginning of the school year from the non-enrolled.

[2]Since the distance in time between the survey and the beginning of the current school year is necessarily less than one year, children can have passed at most one birthday in the meantime.

In formulae: if $p(a_s = a_t) \leq p(e = 1)$, then

$$p(a_s = a_t|e) = \begin{cases} \frac{p(a_s=a_t)}{p(e=1)} & e = 1 \\ 0 & e = 0 \end{cases}.$$

If $p(a_s = a_t) \geq p(e = 1)$, then

$$p(a_s = a_t|e) = \begin{cases} 1 & e = 1 \\ 1 - \frac{p(a_s=a_t-1)}{p(e=0)} & e = 0 \end{cases}.$$

This is at the entry age. At the graduation age, in other words, the age of 12 in the above example of entry at 6 years into a 6 year primary cycle, the order is reversed. That is, the assumption is that the enrolled are preferentially assumed to having been 11 years at the beginning of the school year and therefore being counted in the calculation of the enrolment rate for the age range 6–11, while the non-enrolled are assumed to be drawn preferentially from among those who were already outside the age range at the beginning of the school year. In the above formulae, this corresponds to switching $e = 1$ and $e = 0$.

The strictly ordered bound procedure outlined above is an underestimate of the out-of-school rate, because if there are *both* 6-year-olds of exact school age who are not enrolled and 6-year-olds below exact school age who are enrolled, then the correction above will include some 'wrong' children in the corrected enrolment rate, and equivalently for the 12-year-olds.

The uniform-birth-month assumption might underestimate the effect, if the probability of remaining enrolled until the final grade is greater for the relatively older members of each entry cohort. Note that this is not about students who are in fact over-age. The latter may well experience lower chances of progression, if they were late entrants or repeaters. If the final grade is filled disproportionately with relatively older students, the age-related decline in apparent enrolment rate will be steeper, but not deeper.

If there is variation in survey time t, and/or some additional data are available that provide independent information on the conditional probabilities above, it may also be possible to estimate the conditional probabilities through statistical modelling. An example of this is provided in the country case study in Section 4.

If the survey field work was spread out over several months, the above procedure needs to be carried out for each survey time period t, since both the probabilities above depend on the probability $p_t(a_s = a_t)$ which depends on t. This can either be done by applying selection probabilities to individual observations depending on their exact survey time t and bootstrapping, or by aggregating the survey times by month, and applying the probabilities as proportions to the aggregate numbers. While the adjustments to aggregate indicators such as the net enrolment rate or the share of over-age enrolment can be performed on aggregate numbers, the conditional probabilities can also be included in statistical analyses of the determinants of educational participation that proceed at the micro, i.e. individual, level. Instead of adjusting the shares of enrolled and non-enrolled at different ages, the conditional probabilities above can be used to sample populations of assumed true school age from the raw data and perform the analysis on these resamples. Alternatively, if estimation

is performed within a fully Bayesian framework, the conditional probabilities can also be incorporated directly into the model.

4. Applications and Example: Out-Of-School and Over-age Children in Indonesia

The measurement distortions discussed above would be of limited interest if they did not directly and substantially affect several indicators at the core of international educational development debates, such as the number of out-of-school (OOS) children, or the presence of over-age enrolment. The following sections aim to establish both the demonstrable existence of the above distortions in real-life data, and an approach to correcting for them to arrive at adjusted estimates of the above indicators. As a country case study, Indonesia is particularly well suited to analyse these effects, for several reasons.

Indonesia is large in population overall, and in school-age population. Any errors in estimating its educational participation therefore contribute significantly to the global value. There is fairly good data availability, so that *both* administrative and survey data are analysed in UIS (2005), which also selected it as one country for in-depth analysis. Indeed, due to its importance and dynamic educational expansion trajectory, Indonesia is a popular case study for research on educational development (e.g. **?**), and the results here can be seen to be relevant to a wide body of research. The survey in question is an internationally comparable *Demographic and Health Survey* (DHS), and offers the advantage for present expositional purposes that the field work period covers several months and is approximately centred on the mid-point of the school year.

One slight complication concerns the primary school age range. Some sources specify the official school entry age as being 7 years, and accordingly the age range for the primary cycle to be 7–12, based on the fact that the Indonesian constitutions mandates the authorities to provide schooling for this age range. However, as will become evident below, the empirical enrolment behaviour suggests that in practice this is interpreted so as to ensure that children are in school at the time they turn 7 years old. In other words, the normal entry age is 6 years of age. This raises an interesting question exactly what kind of disadvantage the notion of 'over-age' enrolment attempts to capture. There is, after all, nothing inherently good or bad about starting school at age 5, 6, or 7, and the exact same age-grade combination may be over-age in one country but not in another. Entering school at age x is only indicative of some social or institutional malperformance or a predictor of poor outcomes in its relation to some reference age. The question is whether the most meaningful reference age is the official entry age, however theoretical, or the students age relative to his or her actual peers. To this end one may perform thought experiment which of the following two situations should be assumed to indicate more adverse outcomes for an individual: in the first, our student enters school at the official entry age of 7, but almost everyone else actually enters at age 6; in the second, our student enters school at age 7, even though the official age is 6, but so does almost everyone else. Either way, this question will not be pursued further. Since the question of present interest is how, given some definition of over-age enrolment, the value of

this indicator is affected by survey timing, in the following, the age range 6–11 is taken as the reference ages for the sake of argument, corresponding to the *de facto* normal entry age.

Beginning, with the simplest case, where we label students as 'over-age' for the primary cycle if they are enrolled in primary, but older than 11 years. By contrast, 8-year-olds in grade 1 are not counted as 'over-age', for example. In other words, the issue is whether a student is 'over-age in cycle'. The reason this measure is then distorted when integer ages are observed at an observation time after the start of the school year is that, referring back to Figure 2.1, we are incorrectly excluding triangle (d) from the 'properly aged' student body. It may seem as if this issue could be resolved if the threshold for over-age enrolment is defined as being *more than* one year above the norm age for grade. Indeed, such a 'grace period' is commonly granted.

However, the argument that some children observed at 1 year over age were in fact a year younger at the relevant reference time (at the school start) can be translated to higher ages. This also means that some children who are commonly assumed to be part of the '2 years over age' category actually belong in the '1 year over-age' category, and so on. For this reason, in UIS (2005), for example, it would be the estimate for the '3+ yrs over-age' group that will diminish if exact ages are considered, not necessarily, or not only, the '1–2 yrs over-age' category. The latter loses some children who are not, in fact, over-age, but gains others who are, in fact, only somewhat over-age instead of highly over-age.

Iterating this argument shows that the number of those who are many years over-age is likewise reduced (indeed, the relative reduction in their number may well be greater)[3].

[3]Whether the relative reduction is greater for the number who are over-age by few or many years depends on whether the fall-off of over-age students is steeper than geometric. First note that it is only possible to be over-age by an integer number of years. Since a student who enters aged 6.5 is not over-age, on account of not having been eligible for entry the year before (assuming a standard entry at 6 years), a student entering aged 7.5 is 1 year over-age, not 1.5. If the number of entrants n years over-age at time t, call it $f(n,t)$, follows a geometric progression, then for all n

$$\frac{f(n+1,t)}{f(n,t)} = r,$$

for some constant $0 \leq r \geq 1$. If the decline with n is less rapid than geometric, we have

$$\frac{f(n+1,t)}{f(n,t)} > \frac{f(n,t)}{f(n-1,t)}.$$

Now if t_b marks the beginning of the school year and t_e its end, then, since $f(n,t_b) \approx f(n+1,t_e)$ we see that

$$\frac{f(n,t_b)}{f(n,t_e)} > \frac{f(n-1,t_b)}{f(n-1,t_e)},$$

In words, the ratio of the corrected number of those n years over-age to that observed at the end of the school year is greater than the corresponding ratio for $n-1$. Conversely, if the fall-off is more rapid than geometric, the opposite is the case, and for greater n there is a more pronounced reduction in the estimate.

Positive population growth exacerbates this effect. Since those more over-age come from earlier, smaller, birth cohorts in this case, those appearing as n years over-age are recruited from among the members of cohort c at the beginning of the school year, but from cohort $c+1$ at its end.

The data in which the analysis is based are extracted from the Indonesian DHS 2003/04. The weighted sample contains around 20.000 observations of children aged 6–11. With a focus on triangle (e), Figure 4 plots the share attending school among those aged 6 years at the time of survey, by the month the household was interviewed, for rural households. The dashed line indicates the theoretical share of those who had already completed 6 years of age at the beginning of the school year, assuming a uniform birth month distribution. A linear least squares regression line to the observed points is included.

Several observations can be made with respect to this figure. If it were the case that school entry was universal among 6-year-olds at the beginning of the school year, and zero among 5-year-olds, the observations and regression line would, up to sample variation, coincide with the dashed line. The deviations from it arise from multiple factors.

To begin with, survey timing might be correlated with predictors of participation. Since such confounding bias is a nuisance error from the point of view of understanding the demographic shift in the indicator, this is not investigated further here. In fact, the urban/rural differential is a case in point. The share of urban households interviewed in Feb (ca. 7.5 months after the school year) was almost twice as high as during the other months. Given the fact that enrolments are also higher in urban areas, this creates an error that is effectively not independent of survey timing. This is also the reason why the analysis shown is restricted to the rural subsample.

In addition to possible additional confounders correlated with survey timing, the reason the estimated relationship diverges from the theoretical one is structural. Suppose that, at the beginning of the school year, 6-year-olds enter at some rate r_6 but that 5-year-olds also enter at some lower rate r_5, and suppose that these rates depend only on the integer, not the exact age, of the entrants. It is clear that in the diagram, which depicts the attendance of 6-year-olds *at the time of survey* indicated on the x-axis, at $x = 0$ and $x = 12$ the observed attendance would be approximately r_6 and r_5 respectively. At intermediate values of x, the observed attendance would be a mixture of the two, with the weights changing linearly from 1 to 0 and vice versa over the course of the school year.

Reversing this insight and the same assumptions, we can derive the values for r_6 and r_5 that are implied by the observed regression line. While in this particular case the OLS regression line does indeed intersect $x = 0$ and $x = 12$ at feasible values in the interval $[0, 1]$, a general approach needs to perform the estimation subject to this constraint. A simple Bayesian estimation of the observed points as binomial outcomes with a probability moving linearly between points $(0, y_0)$ and $(12, y_{12})$, with uniform priors for y_i on the interval $[0, 1]$ meets this requirement. In the present case, the estimated entry rates (posterior means) are $r_6 \approx .76$ and $r_5 \approx .21$, with 95% credible intervals of $[.67, .84]$ and $[.12, .30]$. These are not *prima facie* implausible. While one-in-five school entry among 5-year-olds seems high, the observed participation even among children who were actually still aged 5 at the time of survey (!) is almost 9%. Indeed, performing the same analysis on children aged 5 at the time of survey gives an estimate for r_5 of 17%, with estimation interval $[.12, .21]$. Given that, in reality, older 5-year-olds are likely to enter school at a higher age than younger 5-year-olds, we would indeed expect to observe a somewhat lower rate among those who were not only aged 5 at the beginning of the school year, but even at the later survey, who

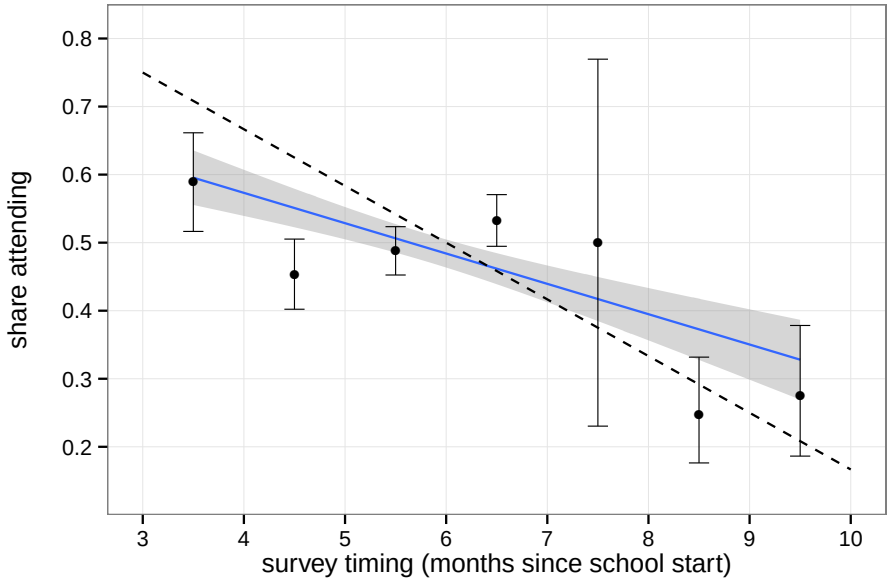

Figure 6: Share attending school among 6-year-olds at time of survey, by survey month. Dot = mean, bars = 95% confidence intervals, solid line = linear regression with corresponding confidence intervals, dashed line = theoretical share who were already 6 years old at the beginning of the school year (i.e. at x=0).

are captured in the second analysis above. As an average rate among 'barely 5-year-olds' and 'almost 6-year-olds', the 21% seem plausible, therefore. In any case, the fact that not only a three-to-one majority of 6-year-olds, but even about one-in-five of 5-year-olds enters school lends further credence to considering the standard entry age to be 6 years rather than 7 in Indonesia.

The results of applying the approach from Section 3 are shown in Figure 4. In addition to the bounds resulting from the assumptions that enrolment is ordered exactly according to exact age, and conversely that enrolment is entirely independent of exact age, a third simulation is run using the above regression to calculate the inverse probabilities that an individual had already completed 6 years of age at the beginning of the school year, conditional on his or her enrolment status at the time of survey. The regression probabilities are only applied for the observed 6-year-olds. For higher ages, the random independence assumption is applied. This is because an equivalent analysis for higher ages would need to carefully consider actual drop-out behaviour, for which we do not have detailed data that is finely grained along time. In any case, the results show that even restricting the regression-based probabilities to just the 6-year-olds is sufficient to significantly change the estimate.

As expected, the adjusted estimates for the share of primary school age children 6–11 *at the beginning of the school year* who are out of school are all lower than the figure based on the 6–11 year olds at the time of survey. The order of the three adjusted estimates also corresponds to what is expected. That is, the lowest estimate results from the assumption of strict ordering of enrolment status by exact age, which removes the greatest number of non-enrolled children who were 6 years at the time of survey, on the assumption that

they were in fact only 5, and therefore not of school age, at the beginning of the school year. The highest adjusted estimates arise from the random estimation of true age at the beginning of the school year. Since these are upper and lower bounds (barring a scenario where 5-year-olds are *more* likely to enrol than 6-year-olds), it comes as no surprise that the empirical regression-based estimate falls in between the above two. In this particular case, it happens to fall almost exactly in the middle. While this does not necessarily generalise in terms of the true value, it suggests that in the absence of the information necessary for a regression-based estimate, the middle of the bounded interval may represent a workable first-order approximation.

Apart from the order of the estimates, we note that the absolute difference between the customary indicator and the adjusted estimates is large enough to potentially affect policy conclusions. This conclusion is confirmed by examining the impact of the adjustment when regression modelling is performed to analyse the determinants of school participation. The estimated effects of a number of standard contextual variables on school enrolment of 6–11 year-olds (at the time of survey in the case of the standard approach, and estimated at the beginning of the school year in the case of the adjusted ones) are displayed in Table 1. What stands out most is the fact that the parameter estimates based on the adjusted-age population show a rather larger wealth differential in terms of enrolment probability. Again, this may potentially lead to different policy conclusions, especially if time series or cross-country comparisons are based on surveys that were conducted at different times in the school year and therefore adjusted by different amounts.

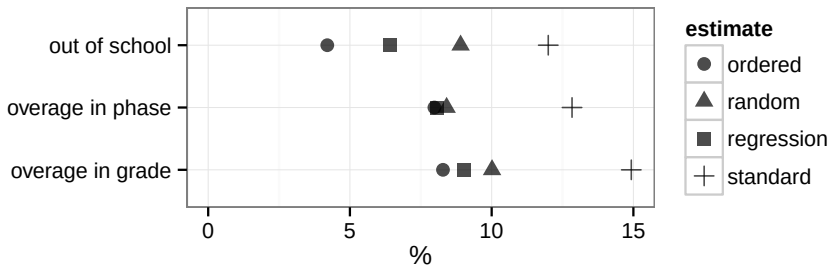

Figure 7: Estimated participation measures for children aged 6–11 in Indonesia, by handling of ages. 'Overage' indicates being 2+ years older than the norm age. 'Standard': no adjustment to discrete ages at time of survey. For all other estimates, only a share of 6-year-olds is retained that corresponds to the timing of the interview. 'Random': exact age assumed unrelated to enrolment status. 'Regression': enrolled share among those aged 6 at the beginning of school year estimated empirically. 'Ordered': exact age of the enrolled assumed higher than exact age of the unenrolled.

Table 1: Estimated marginal multiplicative effects (odds ratio scale) on primary school attendance among children aged 6–11 in Indonesia, by handling of ages. Reference category: male, urban, poorest quintile.

estimate	age	female	rural	poorer	middle	richer	richest
standard	2.17	1.12	0.88	1.82	2.37	3.03	4.50
random	1.56	1.13	0.95	1.87	2.49	3.33	4.97
regression	1.51	1.16	0.92	2.00	2.68	3.63	5.39
ordered	1.26	1.23	0.90	2.28	3.17	4.60	6.57

5. Out-of-school children, or out-of-school *childhood*?

The problem of quantifying non-enrolment among the child population of school age is commonly framed as a question of *counting*: how many out-of-school children are there? The answer, even using exact ages for both the school age thresholds and for the respondents, changes over the course of the school year, due to continuous drop-out. Debates can be had about whether the count at the beginning of the school year is an underestimate, or the count at the end of the school year an overestimate. The idea that, if there is such a thing as the 'true' count, it would be an average over the course of the school year can be taken further to reframe the measurement question. Instead of counting children, we may characterise the question as one of counting (fractional) children-years. In less technical language, the underlying question is one of estimating, not the *number* of OOS children, but the *amount* of out-of-school *childhood*. Conceptually, such a perspective has the advantage that drop-out at the beginning of the school year, which makes the student miss almost the entire year, is counted differently from drop-out at the end, when the student has had almost a whole year additional exposure to school.

Of course, neither full individual educational histories nor aggregate enrolment or attendance by month are normally available (or even collected). This is not, therefore, an impractical proposal to calculate OOS childhood rates instead of estimate counts of OOS children. However, an understanding that, conceptually, approximating OOS childhood is what we are after when OOS children are counted at different points in time, does have practical implications.

For one, it leads to the conclusion that the problem of identifying those children who are expected to return to school isn't one. Some children may drop out of school during the academic year, only to re-enter school at the beginning of the following one, in the same grade. From a perspective of classifying individuals, they are arguably repeaters rather than drop-outs. If we conceive of the estimation of OOS children as the estimation of counts of children in different categories, this raises the problem of whether to include them or not, especially if we do not yet know whether they will re-enrol. By making the status of being out-of-school a function of the individual's past and future, we are adopting a cohort perspective, and this inevitably clashes with the aim of estimating a period indicator, namely the number of OOS children *in a given year*. By contrast, if we conceive of the task as that of estimating OOS child*hood*, in other words, the proportion of person-years lived at school-age that are spent in the non-enrolled state, the future behaviour of potential re-entrants becomes a non-issue. Their time spent out-of-school in year t, which is their contribution to the OOS childhood rate, does not change depending on whether they do or do not return to school the following year, and, as a result, contribute or don't contribute to the OOS childhood rate in *that* year. Another practical implication concerns the timing of data collection. The criticism that school registration data collected at the beginning of the academic year underestimates the number of OOS children continues to apply, because those leaving school at some later point in that academic year, and their person-time spent in the non-enrolled state, are not counted. However, it becomes clear that collecting enrolment information at the latest possible time in the year is also a distortion. True enough, it does

not 'miss' any OOS children, and it does not count children who are in fact enrolled. But it does assign equal weight to those who missed the entire year, and those who dropped out the day before the survey. These two groups spent very different amounts of time in the OOS state, and therefore contributed differently to the OOS childhood rate.

For some policy question, it does, of course, matter whether one child missed ten months of school or whether ten children missed one month each. However, a similar abstraction is actually occurring even with counts of OOS children. Some children experience spells away from school, without necessarily having dropped out entirely. Snapshot surveys will count those who happen to be enrolled and attending at the time of the survey as enrolled, and the others as being OOS. This does not create a bias: if n children spend $x\%$ of the school year away from school, then the survey will count $\frac{nx}{100}$ of them as OOS children, and the total implied contribution to the OOS childhood rate is the same. However, it does demonstrate that even counts of OOS children do not truly relate strictly to the number of affected individuals, but that an abstraction away from individuals and towards person-time rates is effectively implied, even if it is not recognised.

6. Conclusion

Educational participation measures are distorted if age is measured in integer years, and the difference is ignored between age at the beginning of the school year and at the time that school participation is observed. This distortion can be substantial, especially if the number of grades in the schooling phase in question is small, and—naturally—if the distance in time between the survey and the start of the school year is large. In particular, the error induced can be much larger in magnitude than sampling variation, especially when census samples are used. The complex interactions with population growth and —in particular— drop-out preclude the formulation of an rule-of-thumb correction factor that is both simple and accurate.

Nevertheless, even without access to the original micro-data, the unobtainable true value of the participation indicators based on exact age can be approximated given aggregate enrolled/non-enrolled counts by integer age and date of survey. In particular, while complex behavioural modelling may in some cases be possible, it is not necessary, since simple bounds can be established quite easily, by assuming in turn perfect and zero correlation between an individual's unknown exact age and enrolment status. The limited illustrative analysis performed here suggests that in the absence of a robust model, the middle of the bounded interval is a plausible candidate for a serviceable approximation. When faced with an error that can easily reach 10–20 percentage points, a first-order correction should clearly not be abandoned just because some error will remain.

The implications for educational development policy are noteworthy. The number of children out-of-school or over-age tends to be overestimated when integer ages without correction are used—as they customarily are. Biased estimates of these indicators are bad enough, but when time series are assembled from surveys that occurred at different times during the school year, changes in these indicators over time will be distorted too.

Moreover, where the field work for a survey is performed over several months, statistical inferences will be biased if the exact time at which the survey was administered correlates with confounding variables, such as geographical location. Since one of the main reasons for spreading out survey field work over time may be precisely to cover different locales or regions one after the other, this is likely to be the rule rather than the exception.

While administrative school data collected at the beginning of the school year does not suffer from all the effects discussed above (although it still incurs an error in using integer ages when the exact school entry age is, in fact, fractional), the case of Indonesia shows that an analysis along these lines can still uncover information that may be missed otherwise—in this case, that actually a large majority of those aged 6 *at the beginning of the school year* enter school, and that the effective primary school age range is therefore 6–11. More importantly, researchers interested in how school participation is affected by context, or differs between different subgroups, have no choice but to rely on survey, rather than administrative data.

A policy recommendation arising from the above analysis is that, ideally, surveys capturing educational data should be conducted at the very beginning (or the very end) of the school year, or at least not between or spanning several school years. Unfortunately for educationalists, this is unlikely to happen in the case of household surveys that are not primarily education surveys and will, therefore, not prioritise the quality of education statistics over other considerations. Given the fact of enrolment or attendance data collected over the course of a school year or between years, a minimum requirement should be to decide on and communicate unambiguous guidelines on including or excluding prospective entrants during the break, and to routinely ask the two-stage question: "Does [NAME] currently attend school?', with a negative response followed by 'Did [NAME] ever attend during the current school year?'. And while this should in any case go without saying, the exact time of survey information should always be obvious and easily accessible. For surveys specifically designed to collect education data, based on a knowledge of local drop-out behaviour (whether it is seasonal, or related to reaching a certain age, for example), it would be possible to determine a window for collecting enrolment/attendance data that is least biased in terms of approximating the OOS childhood rate. In general, there is surprisingly little educational research on the exact timing of drop-out patterns over the course of the school year. The recommendation that 'additional studies need to be conducted to better understand the issue [of age reporting]' (UIS, 2010, 42) can the emphatically re-iterated here.

Finally, in addition to highlighting the measurement problem, and indicating at least approximate ways of correcting for them, a formal demographic perspective also clarifies what it is that the 'out of school children' indicator actually measures.

References

bildungsserver.de, September 2012. Zur Stichtagsregelung in den Bundesländern. `http://www.bildungsserver.de/innovationsportal/bildungplus.html?artid=846`.

IPUMS, 2011. School attendance: Questionnaire text. `https://international.ipums.org/international-action/variables/SCHOOL/#id2010a`.

Porta, E., Arcia, G., Macdonald, K., Radyakin, S., Lokshin, M., 2011. Assessing Sector Performance and Inequality in Education. World Bank.

UIS, 2005. Children Out of School: Measuring Exclusion From Primary Education. UNESCO Institute for Statistics.

UIS, 2010. Measuring educational participation: Analaysis of data quality and methodology based on ten studies. Technical Paper 4, UNESCO Institute for Statistics.

UNESCO Division of Statistics, 1997. Primary and Secondary Education: Age-Specific Enrolment Ratios by Gender 1960/61 - 1995/96. UNESCO.