Review

autism

Autism
2016, Vol. 20(5) 517–527

DOI: 10.1177/1362361315592378
aut.sagepub.com
SAGE

Estimation of the prevalence of autism spectrum disorder in South Korea, revisited

Peter C Pantelis and Daniel P Kennedy

Abstract

Two-phase designs in epidemiological studies of autism prevalence introduce methodological complications that can severely limit the precision of resulting estimates. If the assumptions used to derive the prevalence estimate are invalid or if the uncertainty surrounding these assumptions is not properly accounted for in the statistical inference procedure, then the point estimate may be inaccurate and the confidence interval may not be a true reflection of the precision of the estimate. We examine these potential pitfalls in the context of a recent high-profile finding by Kim et al. (2011, Prevalence of autism spectrum disorders in a total population sample. *American Journal of Psychiatry* 168: 904–912), who estimated that autism spectrum disorder affects 2.64% of children in a South Korean community. We reconstructed the study's methodology and used Monte Carlo simulations to analyze whether their point estimate and 95% confidence interval (1.91%, 3.37%) were reasonable, given what was known about their screening instrument and sample. We find the original point estimate to be highly assumption-dependent, and after accounting for sources of uncertainty unaccounted for in the original article, we demonstrate that a more reasonable confidence interval would be approximately twice as large as originally reported. We argue that future studies should give serious consideration to the additional sources of uncertainty introduced by a two-phase design, which may easily outstrip any expected gains in efficiency.

Keywords

autism spectrum disorders, epidemiology, prevalence, two-phase screening

Introduction

The secondary reporting of community- and clinic-based diagnoses of autism spectrum disorder (ASD) often lacks the rigor of gold standard evaluations used in scientific research and may be biased by current diagnostic trends, parental concerns, and the practical exigencies of the modern health-care system. Therefore, in order to achieve the best possible estimate of ASD prevalence, one would ideally test the *entire* target population using gold standard diagnostic tools. However, this approach is impracticable for multiple reasons: first, it is generally impossible to access an entire population; second, gold standard diagnostic tools are expensive in terms of time and resources (the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2001) and the Autism Diagnostic Interview–Revised (ADI-R; Lord et al., 1994) require several hours to be administered, after which a diagnosis is rendered by a clinician who has undergone sufficient training to become satisfactorily reliable with these instruments). Testing the entire population with these costly diagnostic tools would be an inefficient way to uncover cases of a rare condition.

Therefore, the researcher settles for an estimate derived from a smaller sample and hopes to generalize from this sample back to the broader population. There are two basic epidemiological approaches to this study design, each with noted strengths and weaknesses. The first is a *single-phase* design, wherein a random subset of the population is selected, and gold standard diagnostic practices are applied to this smaller subset. This overcomes some costs of testing the entire population, yet still lacks efficiency when the condition is rare; to achieve confidence that the true population prevalence falls within an interval of any useful precision, one may need to administer hundreds, if not thousands, of tests on randomly selected individuals. And for double the precision, one must typically test far more than double the individuals—more like 3–4 times as many.

Two-phase designs attempt to overcome these limitations by introducing a preliminary screening phase, in order to focus testing on a (non-random) sample of the

Indiana University, USA

Corresponding author:
Peter C Pantelis, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th Street, Bloomington, IN 47405, USA.
Email: pcpantel@indiana.edu

population that is more likely to have the rare condition. Gold standard diagnostic evaluations can be concentrated on the screen-positive sample, and the number of confirmed cases in this subset of individuals can be used to generalize back to the broader population. Under ideal conditions, introducing a first-phase screener (perhaps one of many available ASD screening questionnaires; see Baron-Cohen et al., 2001; Chandler et al., 2007; Posserud et al., 2006; Ritvo et al., 2011) theoretically allows for an efficient deployment of scarce resources.

A key factor in deriving a prevalence estimate with a two-phase design is how well the chosen screener discriminates between those who truly have a given condition and those who truly do not (McNamee, 2003). For example, if one employs a screener with perfect sensitivity (i.e. 100% of screened individuals with ASD come up positive), then one can be confident that the screener will catch every case of ASD in the target population (i.e. there will be no missed cases lurking among the screen negatives). But if the screener has less than perfect sensitivity at a given cutoff score—as is generally true—then this must be factored into the design and analysis, and the prevalence estimate must be adjusted accordingly. Furthermore, if one is unsure about just *how* sensitive and specific the first screener is at a given threshold, this necessarily limits the precision of the resulting estimate (Erkanli et al., 1997). For these reasons, using any single cutoff for the first-phase screener—and not sampling any participants below that threshold for the second-phase evaluation—hinders the statistical estimation of prevalence (Dunn et al., 1999).

Compounding uncertainty in both single-phase and two-phase designs is that the non-random selection of study participants (e.g. via non-response bias) can greatly affect the results (Posserud et al., 2010). For example, if individuals with ASD were twice as likely to respond to an advertisement to participate in a research study, then the respondents would no longer be a representative sample of the broader population from which they were drawn, and this non-response bias would ultimately result in a gross overestimate of ASD—unless the extent of this bias was quantified and factored into the analyses to appropriately adjust the prevalence estimate.

The optimal choice between a single-phase and two-phase design can strongly depend on the circumstances—for example, the rarity of the condition and the relative cost and performance of the screener compared to full assessment (Shrout and Newman, 1989). However, single-phase designs are more straightforward to administer and analyze (Deming, 1977), and because of the methodological complications that can be introduced by two-phase designs, some thoroughly welcome their demise (Prince, 2003). Considerations like these have ignited an important discussion on the pages of *Autism* about whether some of the higher ASD prevalence estimates reported in recent years should be trusted (Durkin et al., 2014; Mandell and Lecavalier, 2014; Newschaffer, 2015). Here, we develop some of these arguments in a thorough and quantitative manner, illustrating the consequences of particular methodological choices on resulting prevalence estimates. Rather than making purely theoretical points, we use a recent high-profile finding as a case study (Kim et al., 2011).

The authors of this study used a two-phase design to estimate that ASD affects 2.64% of 7- to 12-year-old children in a district of South Korea. This was a striking finding and attracted attention not only because the prevalence was higher than any other published estimate (Elsabbagh et al., 2012), but also because the study concluded that 90% of those with ASD in their target population were attending mainstream schools, and 72% had no history of psychiatric or psychological services whatsoever. The same data continue to be expanded upon in subsequent studies (Kim et al., 2014), and this research group's findings raise important questions. Does a hidden majority of undiagnosed ASD cases also exist in other countries? Is there something unique about South Korea that deserves additional investigation with respect to this unusually high rate?

At the outset, we wish to make it clear that this is not the only study for which these concerns may apply. One reason why we selected this study was because its highly ambitious scope and provocative conclusions resulted in commensurately high impact, via both the scientific literature (more than 450 Google Scholar citations to date) and mainstream channels (e.g. the statistic is listed on the official website of the Centers for Disease Control and Prevention). Another important factor was that this paper contained sufficient, comprehensible methodological detail, such that we could re-explore the analyses and work through them in quantitative ways.

We hope to convey a series of points, both general to the problem of prevalence estimation and specific to this particular case study.

1. The soundness of underlying assumptions is critical for the derivation of a meaningful prevalence estimate.
2. Certain assumptions can ultimately constrain other assumptions. A set of assumptions can be self-falsifying.
3. The uncertainty surrounding assumptions must be accounted for in the estimation procedure; otherwise, confidence intervals will be artificially small.
4. Conclusion: in this particular case, a prevalence estimate of 2.64% was derived from implausible assumptions, resulting in an unreliable estimate and inflated confidence in the precision of this estimate.

Methods

To begin, we summarize the methodology of Kim et al.'s (2011) paper, upon which we base our arguments. For complete details, we refer the interested reader to the original publication and its supplementary materials.

The target population consisted of the 55,266 children (aged7–12 years) living in the Ilsan district of Goyang City, South Korea, at the time of the study. The study's point prevalence estimate of 2.64% implies that, of the children in the target population, $55{,}266 \times 2.64\% = \textit{1459}$ were estimated to have had ASD. As part of the authors' calculation, they estimated that *150* of these children with ASD were on the disability registry or in special education schools. Rates of ASD among this small subset of children ($n = 294$) were considered briefly and separately by the authors and do not factor prominently in the ensuing concerns; nevertheless, this presumed rate implies that $1459 - 150 = \textit{1309}$ children with ASD were estimated by the authors to have been attending regular education schools.

Of the 41 regular education schools in the district, 30 agreed to participate in the study. Not all families in the participating schools consented to participate in the first-phase screener (the Autism Spectrum Screening Questionnaire (ASSQ)). One key assumption the authors used to derive the prevalence estimate (which we will revisit) was that the schools that participated and the self-selected families who ultimately responded to the screener were no more and no less likely to have children with ASD—an explicit assumption of zero non-response bias. See Figure 1 for an illustration of this sampling procedure, presumed to have achieved an essentially random sample of 23,234 out of 54,972 children enrolled in regular schools in this district.

Given the assumed lack of bias in school participation, the authors' best estimate of the number of children with ASD attending the participating regular education schools would be $1309 \times 66.6\% = \textit{871}$. Then, given the assumed lack of self-selection bias in response to the screener, the authors' best estimate of the number of children with ASD whose families consented to participate would be $871 \times 63.5\% = \textit{553}$. At this point, the 23,234 total children were screened (of which 553 should have had ASD, given a prevalence of 2.64%), primarily with the parent-completed portion of the ASSQ (about 1% of these children also had the teacher-completed portion of the ASSQ completed for them). Exactly 1742 children screened positive, and 21,492 screened negative.

Unfortunately, the precise sensitivity and specificity of the ASSQ at the cutoff scores used are unknown, and therefore one could not, at this phase, know just how many of the 1742 screen positives were true or false positives, nor how many of the 21,492 screen negatives were true or false negatives (see Table 1). But making some estimate of the sensitivity and specificity of the first-phase screener is an inescapable and critical step in a two-phase design that, like this one, does not sample any of the screen-negative participants (comprising 93% of respondents) for further assessment at the second phase (Erkanli et al., 1997).

One could use the second phase to *calibrate* the first-phase screener (Bekmetjev et al., 2012); that is, sensitivity and specificity of the screening procedure (i.e. the first phase) could be estimated empirically by testing rates of ASD in both the screen-positive and screen-negative samples with the gold standard diagnostic procedure (i.e. the second phase). Alternately, the sensitivity and specificity of the ASSQ at a particular threshold could be approximated from previous validation studies (Ehlers et al., 1999; Posserud et al., 2009; see Table 3). Or, one could avoid setting a strict cutoff altogether, instead stratifying the sample into score ranges (from the top to the bottom of the scale) and sampling from each stratum for second-phase evaluations.

In this study, the authors take none of these approaches with respect to estimating the sensitivity of their screener, instead simply assuming it to be 100%. This was not the most plausible assumption, per se, but it was the assumption that was—as the authors correctly note—least likely to result in an overestimation of prevalence. We do not scrutinize this assumption at present; rather, we mention it because of its implication: given that 553 children with ASD would have been screened under the assumptions made to this point, all *553* children with ASD would have screened positive under a further assumption of 100% sensitivity.

This implies that the authors assumed 0 false negatives and used the second phase of the study to derive an estimate of 553 true positives (see Table 2). Although a complicated stratification procedure was employed in the second phase to derive this latter estimate of the number of true positives (a procedure that is difficult to faithfully reconstruct without a full complement of methodological details and data beyond the scope of the original publication, and which we thus treat as a "black box"), a simple weigh-back procedure was then used to work backwards from the screening phase to the top of the flow chart, in order to derive the total prevalence estimate for the target population.

These are important bits of information because, given the number of presumed true positives (553), one can infer that there would have been 1189 false positives and 21,492 true negatives (see Table 2). Thus, one can reconstruct the implied specificity of the screening procedure (the percentage of individuals *without* ASD that correctly screened negative): *95%*.

We now ask the following questions: Is the presumed performance of this screener (100% sensitivity and 95% specificity) plausible? Is a point prevalence estimate of 2.64% reasonable? What about an estimated 95% confidence interval of (1.91%, 3.37%)?

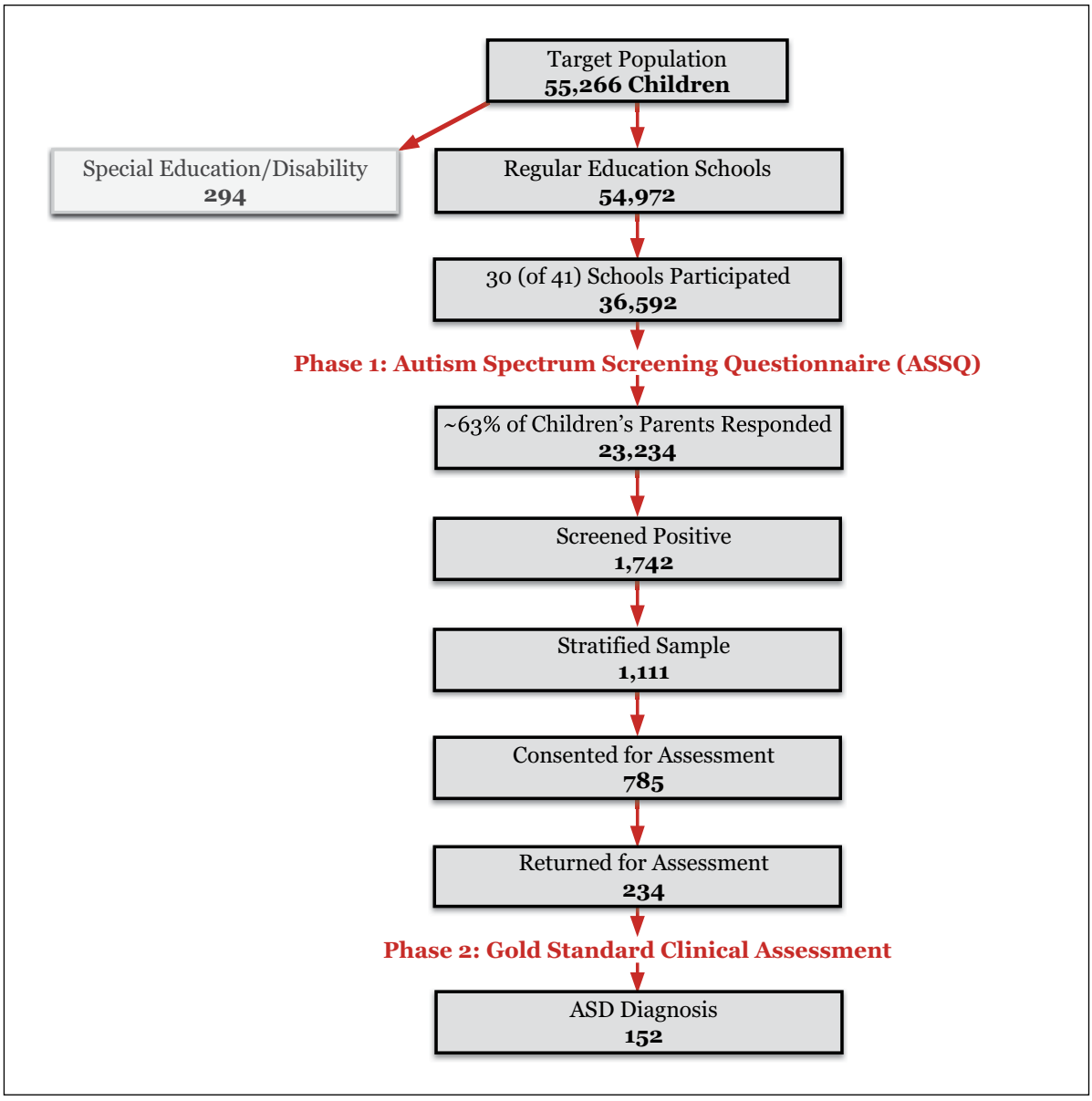

Figure 1. A flow chart of the methodology of Kim et al. (2011). Additional details can be found in the original article.

In order to address these questions, we perform Monte Carlo simulations of Kim et al.'s study. Simulations start at the beginning of the flow chart displayed in Figure 1 and mirror the methodology outlined above under various starting assumptions. Details of this simulation procedure are provided in Supplement 1 (provided online).

Results

The soundness of underlying assumptions is critical for the derivation of a meaningful prevalence estimate

Above, we identified several steps where assumptions were introduced by the authors in order to derive subsequent values. Specifically, assumptions regarding non-response bias (assumed to be zero) and sensitivity of the screener (assumed to be 100%) appear to be critical. What would be the implications of choosing different starting assumptions?

For example, the authors acknowledge the stigma around mental health diagnoses in South Korea (see Kang-Yi et al., 2013). So, what if, due to this stigma, the parents of children with ASD were actually 30% *less* likely to fill out the screener than previously assumed (corresponding to a 44% ASD participation rate instead of 63.5%)? This would increase the point prevalence estimate to 3.7%—a 40% increase in the resulting estimate results from a smaller change to a single underlying assumption. Alternatively, what if one assumed that the parents of children with ASD were 30% *more* likely to participate than

Table 1. The four possible outcomes of a screener like the ASSQ: true positives, false positives, true negatives, and false negatives.

	Screen positive	Screen negative	Total
Children with ASD	True positives?	False negatives?	?
Children without ASD	False positives?	True negatives?	?
Total	1742	21,492	23,234

ASSQ: Autism Spectrum Screening Questionnaire; ASD: autism spectrum disorder.

Table 2. The true positives, false positives, true negatives, and false negatives yielded by the ASSQ, as implied by Kim et al.

	Screen positive	Screen negative	Total
Children with ASD	553	0	553
Children without ASD	1189	21,492	22,681
Total	1742	21,492	23,234

ASSQ: Autism Spectrum Screening Questionnaire; ASD: autism spectrum disorder.

Table 3. A brief review of ASD screening instruments.

Autism screening tool	Study	Samples	Sensitivity	*d′*
ASSQ (Swedish; parent)	Ehlers et al. (1999)	ASD+ vs learning disability	⩽48%	1.4–2.1
ASSQ (Swedish; teacher)	Ehlers et al. (1999)	ASD+ vs learning disability	65%	1.3–1.9
ASSQ (Norwegian; parent)	Posserud et al. (2009)	Total population	⩽91%	⩽2.1
ASSQ (Norwegian; teacher)	Posserud et al. (2009)	Total population	⩽83%	⩽2.1
ASSQ (Norwegian; Max[Teacher, Parent])	Posserud et al. (2009)	Total population	⩽91%	⩽2.4
ASSQ (Chinese; Parent)	Guo et al. (2011)	ASD+ vs. ASD−	⩽82%	2.3–2.9
Autism-spectrum quotient	Baron-Cohen et al. (2001)	ASD+ vs general population	86%–95%	2.4–3.0
Social Responsiveness Scale	Constantino et al. (2007)	PDD+ vs PDD−	75%	2.4
Social Communication Questionnaire	Chandler et al. (2007)	ASD+ vs general population	88%	2.8

Sensitivity arbitrarily depends on the cutoff selected; thus, we provide the (approximate) sensitivity of each instrument that would correspond to the 95% specificity implied by Kim et al. By contrast, *d′* is theoretically threshold-independent and, here, conveys how well an instrument discriminates individuals with ASD from individuals without ASD.
ASSQ: Autism Spectrum Screening Questionnaire; ASD: autism spectrum disorder; PDD: pervasive developmental disorders.

previously assumed (corresponding to an 83% ASD participation rate instead of 63.5%)? The point prevalence estimate would drop to 2.1%.

Similarly, what if the authors assumed a first-phase screener sensitivity of 80%, instead of 100%? The point prevalence estimate would increase to 3.2%.

Thus, the choice of assumptions heavily influences the derived estimate. Without additional information as to which values make the most sense (in terms of either non-response bias or sensitivity), confidence in the resultant estimate will necessarily be limited.

Assumptions constrain other assumptions

As we have shown above, assumptions factor in quite heavily into the calculation of a prevalence estimate. Here, we show that assumptions can implicitly constrain the possible values of other key parameters entering into the prevalence calculation and that *a set of assumptions can indeed be self-falsifying.*

Because only 7.5% of children (1742 out of 23,234) screened positive at the first phase, the specificity of the screening procedure was actually constrained to be quite high; in fact, no matter what the presumed underlying prevalence, non-response bias, and screener sensitivity, the specificity of the screening instrument would have had to fall within a narrow range (93%–96%). If 2.64% prevalence is taken to be true (the point estimate that the authors derive), and zero non-response bias and 100% screener sensitivity are also assumed (the assumptions used by the authors to solve for that point estimate), then this set of constraints further implies that the specificity of the screener would have had to fall within a still narrower range (94%–95%). This specificity value (approximately 95%) was not *explicitly* assumed or provided to the readers, but was constrained by the other assumed and derived values. Now, we ask whether a sensitivity/specificity combination of 100%/95% is actually characteristic of the screener they employed (the ASSQ).

High sensitivity (such as the assumed 100% screener sensitivity) generally comes at the expense of reduced specificity; in order to capture 100% of individuals with ASD with a non–gold standard instrument, there will be a substantial number of false positives. No screener has perfect sensitivity *and* specificity; otherwise, it would by definition be considered the gold standard.

Any screening instrument can be made to be 100% sensitive by imposing an arbitrarily low cutoff score. In the extreme case, a screening instrument that identifies everyone as having a condition would not only be 100% sensitive (i.e. producing no false negatives) but also be 0% specific, and therefore wholly ineffective. Sensitivity and specificity are directly related to one another and cannot be considered separately.

To facilitate comparisons across studies, we convert the sensitivity/specificity combination into a single d' value,[1] which is a single measure of how well a screener discriminates between those with or without a condition. Table 3 compiles the performance of the ASSQ across various studies, alongside the performances of some other commonly used screening instruments, for additional context. When the ASSQ was previously used with a total population sample (as in this study), d' was (at best) 2.4. Yet, the presumed d' of the ASSQ in this study was at least 4,[2] far more discriminative than the best performance reported in any previous study,[3] including the studies that developed the instrument (Ehlers et al., 1999; Posserud et al., 2009).

We argue that this previously reported d' of 2.4 is likely the ceiling performance for the ASSQ, as applied in this article, because of numerous practical challenges faced by Kim et al.

1. The optimal threshold derived by Posserud et al. (2009) was set with retrospective knowledge of the positive and negative ASD cases, and thus was susceptible to overfitting. The threshold used by Kim et al. was set a priori.
2. Posserud et al. (2009) achieved their best results when using the parent and teacher portions of the ASSQ in tandem. In this study, only a small minority of students (1%) had the teacher portion of the ASSQ completed. The vast majority of screen-positive cases (87%) and screen-negative cases (99%+) were decided on the basis of parent ASSQ alone.
3. A total of 6% of parent respondents did not complete the second page of the ASSQ; thus, many of the items were inferred for their children.
4. The authors translated the ASSQ from the language in which it was validated to Korean, which seems unlikely to have improved its performance.

These practical constraints were no fault of the authors; indeed, they are consequences of very reasonable and defensible decisions made at various stages of the research process. Nevertheless, each consideration serves to limit the possible performance of the ASSQ in this context. And, given a more realistic estimate of the performance of the screener, this particular combination of prevalence (2.64%), non-response bias (zero), and screener sensitivity (100%) is incompatible.

If the study had assumed a more realistic, lower sensitivity, this would have resulted in an even higher prevalence estimate. That is, an estimate that was already an outlier with respect to the findings of other recent studies would have become even more extreme.

The uncertainty surrounding assumptions must be accounted for in the estimation procedure; otherwise, confidence intervals will be artificially small

It is clearly possible that assumptions of zero non-response bias and 100% screener sensitivity were incorrect. And an acknowledgment that one or both assumptions were possibly incorrect is the same as an acknowledgment that these assumed values could not have been known with perfect certainty.

A lack of certainty regarding assumptions is not a fatal problem, as long as researchers acknowledge these assumed parameters as meaningful sources of uncertainty surrounding the estimate they ultimately derive. However, Kim et al. apparently constructed their confidence intervals under the implicit assumption of *precisely* zero non-response bias and *precisely* 100% sensitivity. This would mean that their reported confidence intervals reflected uncertainty resulting from random sampling under the stated assumptions, and uncertainty surrounding the number of true positives estimated at the second phase, but did *not* reflect the additionally overlaid uncertainty surrounding the assumptions themselves. In other words, these values would have entered into the statistical model (or in our case, the simulation) as singular, precise parameters, instead of as a distribution of plausible values for these parameters.

In order for the final derived prevalence estimate to be a meaningful communication to the clinical world, the estimate (and its confidence intervals) should account for all known, important sources of uncertainty. Otherwise, the conclusion would need to be continually qualified (i.e. ASD prevalence likely falls between 1.91% and 3.37% *if* one assumes zero non-response bias *and* a perfectly sensitive screener).

The authors assumed zero non-response bias because there were conflicting reasons why one might reasonably expect bias in either direction (e.g. the conflicting motivations of parents to avoid stigma, vs to obtain information that might help their children). Even so, one still must recognize that this bias parameter is unknown and could therefore—in the absence of any additional information—plausibly exist along a wide range of possible values. The only hard constraints are that any individual could not

have been more than 100% likely to participate in the first-phase screener, that 23,234 individuals actually participated, that 1742 children screened positive, and that 152 screen-positive individuals in regular schools were later confirmed to be true cases of ASD.

Thus, rather than assuming with perfect confidence that 63.5% of children with ASD participated in the screener (i.e. at the exact same rate as children without ASD; an assumption reflected in rows 1, 2, and 4 of Figure 2), it would perhaps be more realistic to assume that zero bias was the most likely scenario, while at the same time acknowledging that respondent rates between ~40% and ~80% were also plausible (a different starting assumption reflected in rows 3 and 5 of Figure 2).

Similarly, one might permit wide ranges of possible screener sensitivities, based on the screener's most likely psychometric properties. In our simulations, we allow for a variety of possibilities, with perhaps the most plausible case being something like 80%—that is, quite good in tandem with 95% specificity, but by no means perfect (a starting assumption reflected in rows 4 and 5 of Figure 2).

We are not the first to make these general points, even about this paper (Charman, 2011; Newschaffer, 2015). However, here we show quantitatively that if one acknowledges that the authors did not know with certainty either the magnitude of non-response bias or the sensitivity of the screener, then the final estimate came packaged within a confidence interval that was artificially narrow in precision.

Row 1 of Figure 2 reflects the confidence interval that could hypothetically result from perfect knowledge of all three unknown parameters introduced by the two-phase design: the probability of a child with ASD participating in the study (first column), the sensitivity of the first-phase screener (second column), and the number of true positives captured by the screener (third column). This situation would result in correspondingly low uncertainty surrounding the final estimate; the narrow 95% confidence interval presented in the fourth column would only reflect random sampling error, and to achieve it would require a gold standard diagnostic test of all 1742 children captured by the first-phase screener, combined with perfect confidence in the other critical assumptions.

Row 2 of Figure 2 represents our model of the authors' estimation procedure. The probability of a child with ASD participating is assumed to be 63.5% (i.e. precisely the same as the non-ASD probability). The screener sensitivity is assumed to be precisely 100%. We model uncertainty around the estimated number of true positives as being normally distributed ($\mu = 553$, $\sigma = 82$; the estimated number of true positives and the uncertainty surrounding this estimate were derived by the authors via a stratification procedure and statistical inference method that we cannot faithfully reconstruct from the details provided in the original article; for the purpose of simulation, we therefore treat the number of true positives as a value arising from an irreducible "black box"). Running simulations under this combination of assumptions replicates the point estimate (2.6%) and 95% confidence interval (1.9%, 3.4%) reported by the authors.

What if one allowed for uncertainty around non-response bias (which we model as being sampled from a beta distribution; $\alpha = 10$, $\beta = 5.75$), while still assuming perfect knowledge of the screener sensitivity? Row 3 shows the 95% confidence interval that results from this modified prior assumption. It expands to (1.7%, 4.3%)—75% larger than that which was reported by the authors.

What if one instead assumed perfect certainty around the non-response bias, but allowed for a range of plausible screener sensitivities (which we model as being sampled from a beta distribution; $\alpha = 12$, $\beta = 3$)? Row 4 represents this modified set of assumptions and expands the 95% confidence interval to (2.2%, 4.5%)—50% larger than what was reported by the authors.

What if one allowed both of these assumptions to reflect a realistic level of uncertainty (row 5)? We argue that a 95% confidence interval reflecting these two sources of uncertainty not accounted for in the original publication would probably resemble (2.0%, 5.4%)—more than twice as wide as that which was originally reported. The point estimate, too, would rise to ~3.3%—owing mostly to the entertained possibility of less-than-perfect screener sensitivity.

Conclusion: A prevalence estimate of 2.64% was derived based on incorrect assumptions and likely presented within a confidence interval that was unrealistically narrow

We have now demonstrated that at least one stated assumption was probably incorrect. We have also shown that the reported confidence interval was artificially narrow, owing to a failure to account for the fallibility of the assumptions.

Here, it should be emphasized that an estimated number of true positives (true cases of ASD who screened positive at the first phase) may actually be consistent with *many* possible underlying prevalences, non-response biases, and first-phase screener sensitivities. Even if one knew that precisely 553 true positives had been captured by the first-phase screener, one would still not have solved for a point, but for a surface (mathematically, a manifold with 2 degrees of freedom) embedded in the three-dimensional parameter space (represented in black in Figure 3). The authors made assumptions to artificially constrain two of these parameters (non-response bias and screener sensitivity) to derive the point estimate (represented by the red circle in Figure 3). This estimate represents the point in the three-dimensional parameter space presented by the authors as the most plausible: 2.64% prevalence, 0 participation bias, 100% sensitivity/95% specificity of the screener. But there was a whole

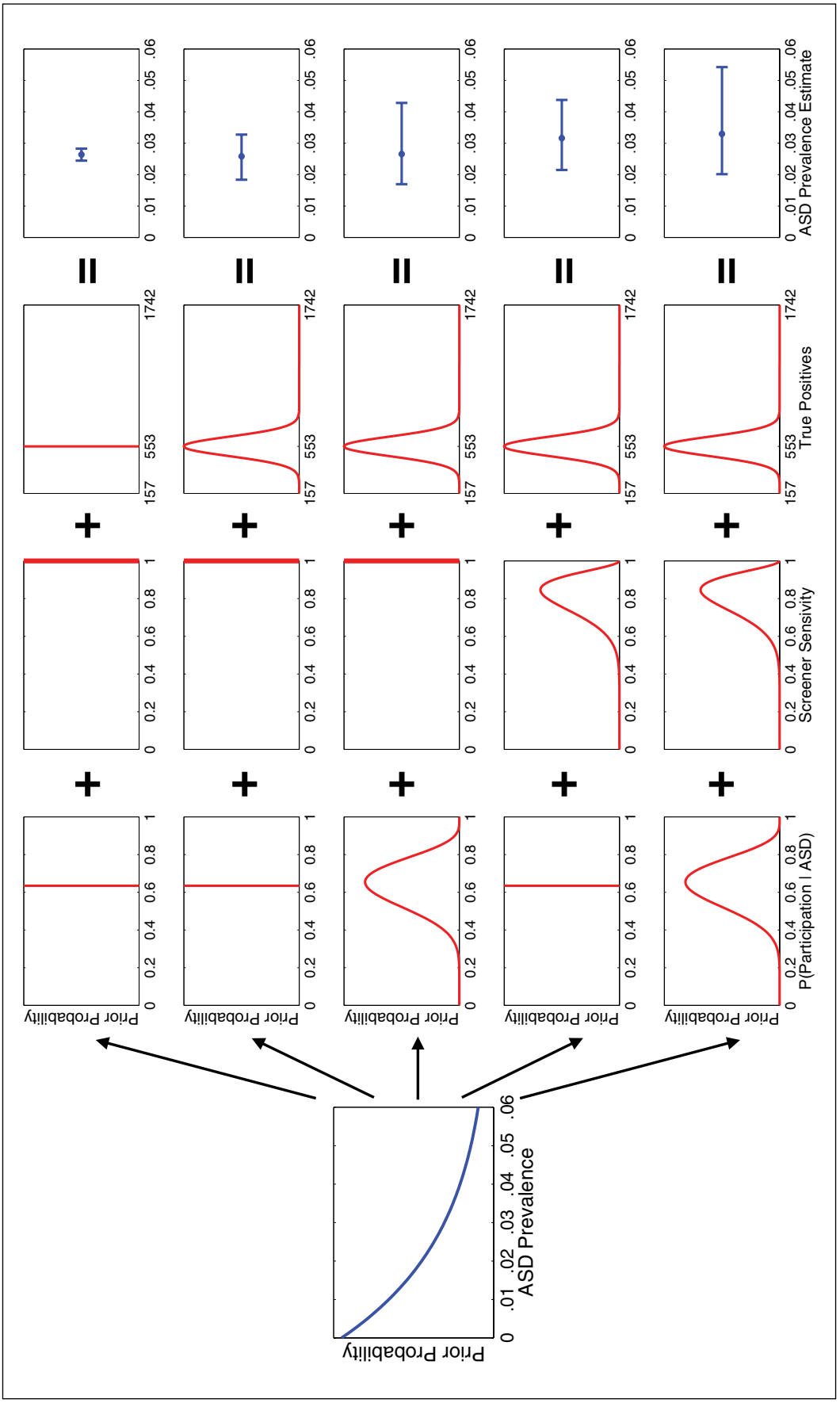

Figure 2. An illustration of how various degrees of uncertainty in prior assumptions can result in prevalence estimates of vastly different degrees of precision. One starts at the left with great uncertainty regarding the prevalence of ASD in the target population (the prior belief is modeled here as a beta distribution; $\alpha = 1$, $\beta = 38$). Each row then represents a different set of starting assumptions. The final column shows the 95% confidence intervals that would result from statistical inferences made under these assumptions. See Supplement 1 for further details about the simulation-based statistical procedure that was used to generate these results.

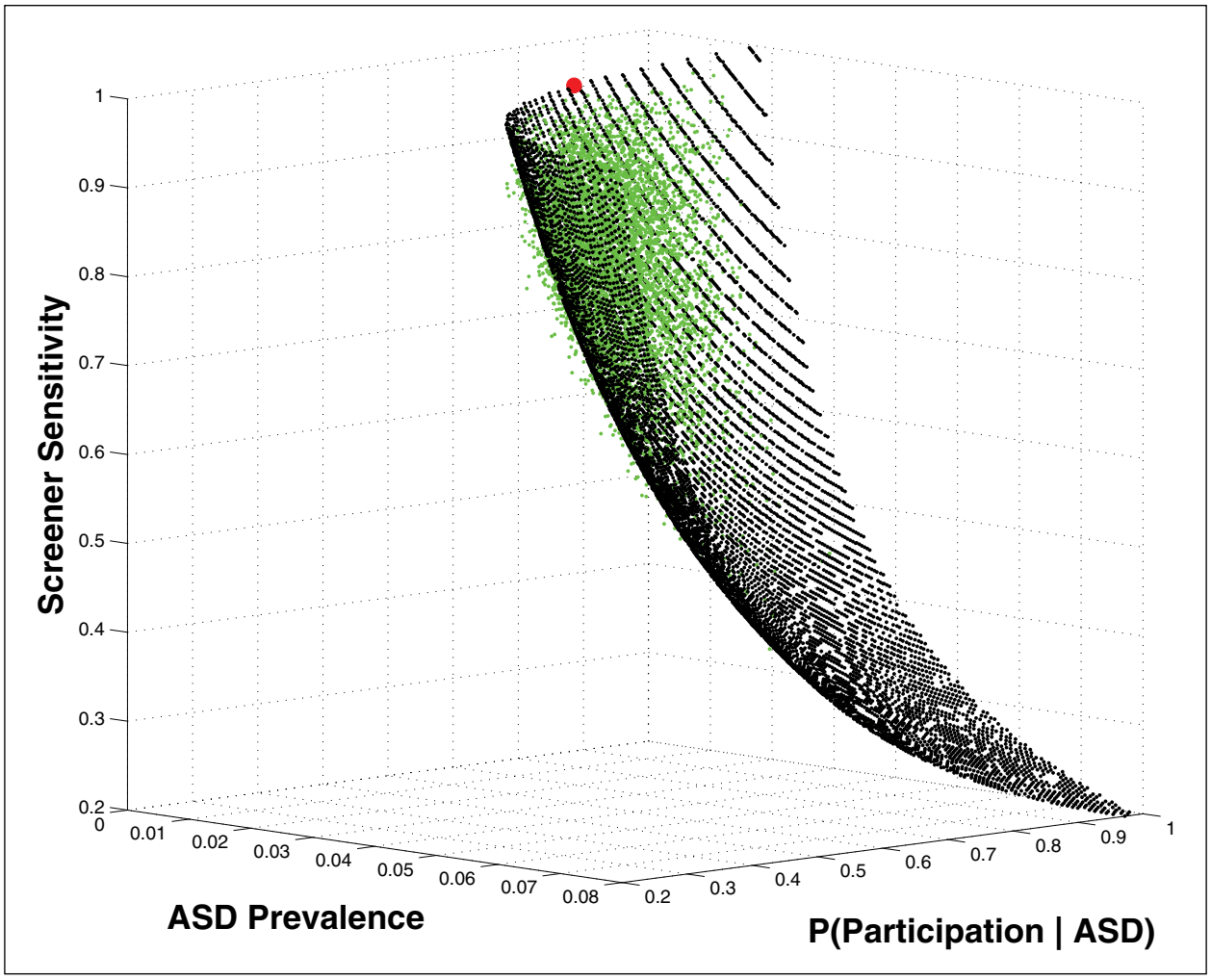

Figure 3. The black surface represents the possible combinations of ASD prevalences, screener sensitivities, and non-response biases that would be consistent with 553 screener true positives. The cloud of green points is sampled in proportion to the plausibility of that combination, given 553 screener true positives. The red circle conveys the point estimate put forth by Kim et al.: 2.64% ASD prevalence, 100% screener sensitivity, and zero non-response bias. (For interpretation of the references to color in this figure legend, the reader is referred to the online version of this article.)

universe of other possibilities to choose from, many of which may in fact have been more reasonable.

Discussion

Two-phase epidemiological studies aim to increase the efficiency by which precise and accurate prevalence estimates can be achieved, relative to a single-phase procedure. By employing a two-phase design, Kim et al. produced an estimate that implied *292* gold standard clinical assessments had been leveraged into a 95% confidence interval for ASD[4] prevalence of (1.91%, 3.37%), purportedly generalizable to a broader population of 55,266 South Korean children. To achieve an estimate of similar precision with a single-phase design (given a true underlying ASD prevalence of 2.64%), one would have had to administer gold standard evaluations to approximately *2000* randomly selected children from the target population—nearly *seven times* as many.

We argue that this level of reported precision is illusory and only achieved by not accounting for sources of uncertainty arising from the introduction of the screening instrument and potential non-response bias. A truer reflection of the precision they achieved would have been conveyed by a point estimate of 3.3% and a confidence interval of (2.0%, 5.4%)—a prevalence estimate that is simultaneously both imprecise and 25% higher than their original estimate, which was already the highest on the scientific record. We emphasize that we do *not* claim that 3.3% would be a more accurate estimate of ASD prevalence; rather, we note that an estimation procedure yielding this result would give us pause to question whether the added methodological and statistical complications introduced by a multiple-phase design had efficiently increased the accuracy or precision of the prevalence estimate in this case.

The authors apparently took great care in the quality control of their gold standard diagnostic procedure, which was especially commendable in light of potentially challenging cultural considerations. The benefits yielded by such care, however, can be hijacked by vulnerabilities elsewhere in a study's design. A one-phase design may be ostensibly inefficient, but will only introduce non-response bias once. Each additional phase introduces new vulnerabilities, and in the case of Kim et al., potential non-response bias was introduced no fewer than three times: once for response to the first-phase screener (which we attempted to account for in our analysis), once in obtaining consent for the second-phase diagnostic procedure, and once when only a minority subset of the consenting individuals actually returned for evaluation. The assumption of precisely zero non-response bias becomes increasingly tenuous as each new opportunity for bias is introduced. The relative simplicity and transparency of a single-phase study—both in terms of administration and subsequent statistical inference—should not be underestimated as practical benefits of this approach.

For researchers embarking on the ambitious task of prevalence estimation in a total population sample, we endorse several prescriptions already put forth by others. The larger the study participation rate, the less possible influence non-response bias can have on the resulting estimate. Before embarking on a two-phase design, one should give serious consideration to the sensitivity and specificity of the screening tools at one's disposal. If one decides upon a two-phase design, then it is probably best to avoid strict cutoff scores altogether, but if such a threshold is applied, then the first-phase screener should be calibrated by the second-phase (gold standard) assessment—which means also testing some of the screen-negative cases. Finally, one should fully explore the implications of all assumptions that factor into the calculation of the prevalence estimate (e.g. how assuming a particular sensitivity would constrain possible specificities).

On the other hand, we acknowledge that statistical efficiency in deriving a prevalence estimate is not always the sole consideration in a study's design. If Kim et al. had assessed a certain number of screen-negative cases with the gold standard diagnostic procedure (as we prescribe above), then limited resources would have been diverted from the gold standard assessment of an equal number of screen-positive cases. Because the latter assessments are more likely to uncover actual cases of ASD and because helpful interventions may be available for the newly diagnosed, ethical considerations at the individual level here trade off with statistical considerations. Kim and colleagues identified many individuals in mainstream schools whose condition was previously unknown to both parents and schools ($n = 152$, corresponding to 0.3% of the population), and the benefits to these children and their families may be quite significant.

As a final note, we do not wish to detract from the much-needed attention the study has brought to mental illness (and ASD in particular) in South Korea and beyond. We remain hopeful that continued research aiming to apply rigorous diagnostic procedures to total populations (like Kim et al., 2011) may indeed result in better estimates of ASD prevalence than have been obtainable in the past.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. This formula may be used to estimate *d′*: *norminv* (sensitivity) − *norminv*(1 − specificity), where *norminv* is the inverse cumulative distribution function of the standard normal.
2. *d′* is infinite when sensitivity is 100%; an estimated *d′* of 4 would result from instead assuming 99% sensitivity.
3. One study (Mattila et al., 2012) that employed the Autism Spectrum Screening Questionnaire (ASSQ) reported a *d′* of 4+, but this study, too, assumed 100% sensitivity at a given cutoff rather than measuring it.
4. The original article reported 286 assessments, adjusted in subsequent *Corrections*.

References

Baron-Cohen S, Wheelwright S, Skinner R, et al. (2001) The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders* 31(1): 5–17.

Bekmetjev A, VanBruggen D, McLellan B, et al. (2012) The cost-effectiveness of reclassification sampling for prevalence estimation. *PLoS ONE* 7(2): 1–6.

Chandler S, Charman T, Baird G, et al. (2007) Validation of the social communication questionnaire in a population cohort of children with autism spectrum disorders. *Journal of the American Academy of Child and Adolescent Psychiatry* 46(10): 1324–1331.

Charman T (2011) The highs and lows of counting autism. *American Journal of Psychiatry* 168(9): 873–875.

Constantino JN, Lavesser PD, Zhang Y, et al. (2007) Rapid quantitative assessment of autistic social impairment by classroom teachers. *Journal of the American Academy of Child and Adolescent Psychiatry* 46(12): 1668–1676.

Deming WE (1977) An essay on screening, or on two-phase sampling, applied to surveys of a community. *International Statistical Review* 45(1): 29–37.

Dunn G, Pickles A, Tansella M, et al. (1999) Two-phase epidemiological surveys in psychiatric research. *British Journal of Psychiatry* 174: 95–100.

Durkin MS, Bilder DA, Pettygrove S, et al. (2014) The validity and usefulness of public health surveillance of autism spectrum disorder. *Autism* 19(1): 118–119.

Ehlers S, Gillberg C and Wing L (1999) A screening questionnaire for Asperger syndrome and other high-functioning autism spectrum disorders in school age children. *Journal of Autism and Developmental Disorders* 29(2): 129–141.

Elsabbagh M, Divan G, Koh Y, et al. (2012) Global prevalence of autism and other pervasive developmental disorders. *Autism Research* 5(3): 160–179.

Erkanli A, Soyer R and Stangl D (1997) Bayesian inference in two-phase prevalence studies. *Statistics in Medicine* 16: 1121–1133.

Guo Y, Tang Y, Rice C, et al. (2011) Validation of the autism spectrum screening questionnaire, Mandarin Chinese version (CH-ASSQ) in Beijing, China. *Autism* 15(6): 713–727.

Kang-Yi CD, Grinker RR and Mandell DS (2013) Korean culture and autism spectrum disorders. *Journal of Autism and Developmental Disorders* 43(3): 503–520.

Kim YS, Fombonne E, Koh Y, et al. (2014) A comparison of *DSM-IV* pervasive developmental disorder and *DSM-5* autism spectrum disorder prevalence in an epidemiologic sample. *Journal of the American Academy of Child and Adolescent Psychiatry* 53(5): 500–508.

Kim YS, Leventhal BL, Koh Y, et al. (2011) Prevalence of autism spectrum disorders in a total population sample. *American Journal of Psychiatry* 168: 904–912.

Lord C, Rutter M and Le Couteur A (1994) Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders* 24(5): 659–685.

Lord C, Risi S, Lambrecht L, et al. (2000) The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders* 30(3): 205–223.

McNamee R (2003) Efficiency of two-phase designs for prevalence estimation. *International Journal of Epidemiology* 32(6): 1072–1078.

Mandell D and Lecavalier L (2014) Should we believe the Centers for Disease Control and Prevention's autism spectrum disorder prevalence estimates? *Autism* 18(5): 482–485.

Mattila ML, Jussila K, Linna SL, et al. (2012) Validation of the Finnish autism spectrum screening questionnaire (ASSQ) for clinical settings and total population screening. *Journal of Autism and Developmental Disorders* 42(10): 2162–2180.

Newschaffer CJ (2015) Regarding Mandell and Lecavalier's editorial "Should we believe the Centers for Disease Control and Prevention's autism spectrum disorders prevalence estimates" and subsequent exchange with Durkin et al. *Autism*. Epub ahead of print 18 February. DOI: 10.1177/1362361314562617.

Posserud M, Lundervold AJ and Gillberg C (2006) Autistic features in a total population of 7–9-year-old children assessed by the ASSQ (Autism Spectrum Screening Questionnaire). *Journal of Child Psychology and Psychiatry* 47(2): 167–175.

Posserud M, Lundervold AJ and Gillberg C (2009) Validation of the autism spectrum screening questionnaire in a total population sample. *Journal of Autism and Developmental Disorders* 39(1): 126–134.

Posserud M, Lundervold AJ, Lie SA, et al. (2010) The prevalence of autism spectrum disorders: impact of diagnostic instrument and non-response bias. *Social Psychiatry and Psychiatric Epidemiology* 45(3): 319–327.

Prince M (2003) Commentary: two-phase surveys. A death is announced; no flowers please. *International Journal of Epidemiology* 32(6): 1078–1080.

Ritvo RA, Ritvo ER, Guthrie D, et al. (2011) The Ritvo autism Asperger diagnostic scale-revised (RAADS-R): a scale to assist the diagnosis of autism spectrum disorder in adults: an international validation study. *Journal of Autism and Developmental Disorders* 41(8): 1076–1089.

Shrout PE and Newman SC (1989) Design of two-phase prevalence surveys of rare disorders. *Biometrics* 45(2): 549–555.