Why Otitis Media is Good for the Language Acquisition: Some Examples of Statistical Pitfalls in Language Assessment Programs

Eugen Zaretsky, Marburg University Hospital, Germany

The Asian Conference on Education & International Development 2026
Official Conference Proceedings

Abstract

Results of statistical calculations can vary depending on the sample size, comparability of subsamples, choice of statistical methods, coding of variables, use of imputation, and many other factors. It is up to the researcher whether the Bonferroni adjustment of *p*-values, or Bonferroni-Holm adjustment, or no adjustment at all is used, whether he or she utilizes parametrical or more conservative non-parametrical statistical methods, whether metrical data are *z*-transformed etc. This presentation shows some examples of statistical pitfalls in large-scale language assessment programs. First, data from German language screening programs are utilized to show why children with frequent otitis media seem to acquire German more quickly than children without frequent otitis media ($N = 1{,}628$). Another example shows why children born in January-July speak better German than those born in August-December ($N = 4{,}316$) and how one can reverse these results on the basis of another sample (the same age, the same region, the same language test; $N = 6{,}144$). One more example demonstrates how a link between stuttering and children's German language competence can be described as very close or non-existent depending on chosen statistical methods ($N = 748$). A high variability of results, in spite of very large samples, validated and standardized language tests, high quality of data in terms of the inter-rater and intra-rater reliability, can be traced back either to errors in the study design or to unexpected confounding variables. For instance, children with frequent otitis media acquired German under comparatively favourable sociodemographic conditions.

Keywords: statistics, language acquisition, language assessment, language screening, German language

iafor
The International Academic Forum
www.iafor.org

Introduction

Results of statistical calculations can vary depending on a number of factors: sample size, comparability of subsamples, coding of variables, use of imputation, use of the Bonferroni adjustment of *p*-values, or Bonferroni-Holm adjustment, or no adjustment at all, use of parametrical or more conservative non-parametrical statistical methods, *z*-transformation of metrical variables or no transformation etc. For instance, too low sample sizes often result in not statistically significant *p*-values (in spite of considerable effect sizes), whereas very large samples can make even smallest differences between subgroups statistically significant. Comparability means that the subgroups that are to be compared should be matched as far as possible regarding well-known influencing variables. For instance, if we compare reading habits of British and Spanish adults, it would be important to match them not only for age and biological sex (or gender), but also for the educational level and occupation because these factors influence reading habits. The term "imputation" refers to a number of statistical methods that fill gaps in data sets to increase sample sizes and make them more representative. For instance, missing values in language tests (no answer) can be filled by mean or median values of other answers. It allows to include in calculations children with a low compliance, that is, those who did not answer some questions in the test. But different methods of imputation deliver different results and we do not know which one can be considered more correct. Also, various methods of adjustment of *p*-values can be used in case of the multiple testing (e.g., five *T*-tests with five subtests of a language screening conducted on the same day with the same child). Use or non-use of such adjustments can influence results considerably, many relevant statistically significant results can be lost due to rigorous adjustments. Furthermore, it is up to the researcher whether he or she chooses parametrical or non-parametrical statistical methods. The latter are more conservative and should be used when certain criteria for parametrical methods are not met (first and foremost, normal distribution of metrical data). In reality, many researchers always use parametrical methods because these deliver more often statistically significant results. Usually, reviewers of manuscripts do not question the choice of parametrical methods and consider them a kind of default option, although, strictly speaking, researchers have to provide evidence that respective criteria were checked (e.g., results of the Kolmogorov-Smirnov test). *Z*-transformation is important when, among other things, two variables with different scales should be correlated (e.g., results of a language test ranging from 0 to 20 and results of an IQ-test ranging from 50 to 150). Usually, differences between correlation coefficients of *z*-transformed and not *z*-transformed values are minimal. But if the *p*-value is marginally significant (about $p = .05$), transformation or non-transformation can change much in the results.

There are many other possible sources of errors or misinterpretations in statistical calculations: missing inter-rater or intra-rater reliability, reporting of results of ANOVA for influencing factors, although the ANOVA model itself was not statistically significant (or explained only 5% of variance in the dependent variable, although 70–80% were achieved in other comparable studies), missing correlation interpreted as no association, although not all associations are linear, etc. Also, it is natural for humans to be satisfied with results if they found evidence in favour of their hypothesis and to keep on searching for possible confounding variables and other sources of errors if results did not confirm their hypothesis. Of course, if they kept on searching for the same statistical issues even in case of "favourable" results, they would sometimes also find some confounding variables and other errors that completely distorted their findings.

The current study focuses predominantly on the problem of confounding variables. Several examples of the influence of such variables are discussed on the basis of data from a large-scale language screening program in the German state of Hesse. Also, influence of the choice of statistical methods on results is shown for the language test data.

Methods

Several samples of children from studies on the development of two language tests were re-analysed retrospectively (Table 1). All monolinguals in Table 1 were children speaking German.

Table 1
Characteristics of Samples Used in the Current Study

	S1	S2	S3	S4	S5
N	1,628	1,791	4,316	6,144	748
Boys	843 (51.8%)	923 (51.5%)	2,306 (53.4%)	3,117 (50.7%)	391 (52.3%)
Girls	785 (48.2%)	868 (48.5%)	2,010 (46.6%)	3,027 (49.3%)	357 (47.7%)
Age range	4;0-4;11	4;0-4;11	4;0-5;11	4;0-4;5	3;10-8;3
Age (months)	53.5	53.5	54.2	—	69.9
Monolinguals	739 (45.4%)	785 (43.8%)	2,039 (47.2%)	4,280 (69.7%)	303 (40.5%)
Bi-/multilinguals	889 (54.6%)	1,006 (56.2%)	2,277 (52.8%)	1,864 (30.3%)	445 (59.5%)
Language tests	KiSS.2	KiSS.2	KiSS.2	KiSS.1	AWST-R, S-ENS, ETS

In samples S1–S4, children's German language skills were assessed by the validated, standardized language screening "Kindersprachscreening" in two different versions ("Language screening for children"; KiSS; Holler-Zittlau et al., 2011). It contains (a) subtests on speech comprehension, vocabulary, articulation, grammar, and phonological short-term memory (repetition of non-words and sentences), (b) questionnaires for parents and kindergarten teachers to assess sociodemographic and medical characteristics of children and their families. Predominantly, the second version of this screening was used, KiSS.2. Results of this test can be dichotomized as follows:

- Need of additional educational assistance in acquiring German (a German language course): yes/no,
- Need of additional medical assistance in acquiring German (some language[-related] therapy, usually speech-language therapy): yes/no,
- "Pass/fail", with "fail" meaning that the child needs educational and/or medical assistance in acquiring German.

In S1-S4, all children were tested in German kindergartens.

In the sample S5, three German language tests were used (for full names, see References):

- AWST-R: vocabulary (Kiese-Himmel, 2005),
- S-ENS: phonological awareness (filling gaps in words), phonological short-term memory (repetition of non-words and sentences), articulation (Döpfner et al., 2005),

- ETS: speech comprehension, grammar (Angermaier, 2007).

S-ENS is one of the German school enrolment tests. All three tests were conducted not in kindergartens (in contrast to S1–S4), but in public health departments. Most children were tested during the school enrolment examination.

The first example of possible pitfalls in statistical calculations is dedicated to children with otitis media. It is an inflammation or infection of the middle ear, the space behind the eardrum. Due to otitis media, children do not hear well, which can slow down the pace of the language acquisition (Nittrouer & Lowenstein, 2024). The KiSS.2 questionnaire for parents contains an item "Does the child often have otitis media?" (yes/no). It was up to parents to define what is meant by "often" (since such definitions are subjective, it represents a certain problem for statistical analyses). Sample S1 was used for these calculations. German language skills and sociodemographic characteristics of children with and without frequent otitis media were compared by means of Chi-Square tests, linear-by-linear associations (lbl), or Mann-Whitney *U*-tests depending on the coding of variables. Children with and without frequent otitis media had comparable age in months according to the Mann-Whitney *U*-test ($p > .05$).

The second example is dedicated to children with cases of dyslexia in the family, that is, with a familial predisposition for language disorders (questionnaire for parents: item "Are there cases of problems with reading and writing (dyslexia) in the family?": yes/no). It was shown in previous studies that such children have relatively poor language competence (Caglar-Ryeng et al., 2019). German language skills and sociodemographic characteristics of children with and without cases of dyslexia in the family were compared by the same methods as described above. Here, sample S2 was used. There was no difference in age between children with and without cases of dyslexia in the family.

The third example is dedicated to months of birth: January–July vs. August–December. There is no reason to believe that children born in certain months should have better German language skills. The comparison of children born in the first vs. second half of the year was chosen to show how results can change, although children were tested in the same region by the same language test (KiSS.2). Here, samples S3 and S4 were used. Chronologically, children in S4 were tested several years earlier than children in S3. There were no differences in age between children born in January–July and those born in August–December in S3. For S4, such comparisons were not possible because no data on exact age in months were available.

The fourth example is dedicated to stuttering. Here, sample S5 was used. Stutterers and non-stutterers were compared regarding their German language skills by means of several univariate statistical methods. Stutterers and non-stutterers were of comparable age. Generally, stuttering is very vaguely associated with weaker language skills, that is, the difference between stutterers and non-stutterers is minimal (Ntourou et al., 2013; Zaretsky et al., 2017). For other languages than German, some studies showed that stutterers can even outperform non-stutterers in language skills if sociodemographic conditions of the language acquisition are favourable (e.g., Reilly et al., 2013).

Results

Results on Otitis Media

According to KiSS.2, children with frequent otitis media were less often classified as needing additional educational assistance in acquiring German (38/108 (35.2%) vs. 717/1,518 (47.2%); $\chi^2_{(1)} = 5.9$, $p = .015$) and, therefore, failed less often in KiSS.2 (42/108 (38.9%) vs. 751/1,518 (49.5%); $\chi^2_{(1)} = 4.5$, $p = .033$). Also, in the subjective estimations of German language skills by kindergarten teachers (item "The child speaks German age-appropriately", from 1 "never" to 5 "always") children with frequent otitis media scored higher than children without frequent otitis media: lbl = 4.7, $p = .030$. Thus, at first sight, it can be concluded that frequent otitis media is a factor that helps children in the German language acquisition. However, a comparison of sociodemographic characteristics of children with and without frequent otitis media shows that this conclusion would be wrong because there are some hidden confounding factors behind this finding. Generally, children with frequent otitis media acquire German under more favourable conditions. They…

- were more often monolingual Germans than bi-/multilinguals: $\chi^2_{(1)} = 14.4$, $p < .001$,
- were less often immigrants (definition: the child or at least one of parents immigrated to Germany): $\chi^2_{(1)} = 4.9$, $p = .047$,
- spoke more often only German at home: lbl = 16.5, $p < .001$,
- began earlier to acquire German: lbl = 17.8, $p < .001$,
- their mothers could read and write German better (self-estimation): lbl = 5.5, $p = .019$,
- their mothers ($Z = -3.3$, $p = .001$) and fathers ($Z = -3.7$, $p < .001$) began earlier to acquire or to learn German.

On the other hand, children with frequent otitis media acquired German under several unfavourable medical conditions. Among other things, they…

- had more often regular medicine intake: $\chi^2_{(1)} = 8.9$, $p = .003$,
- had more often head injuries and operations: $\chi^2_{(1)} = 55.4$, $p < .001$,
- had to attend paediatricians more often than other children: $\chi^2_{(1)} = 73.4$, $p < .001$,
- had more often a permanent hearing disorder: $\chi^2_{(1)} = 24.9$, $p < .001$.

Nevertheless, obviously, their monolingual German, non-immigrant background outweighed negative medical factors and helped children with frequent otitis media in the German language acquisition. Therefore, comparisons of children with and without frequent otitis media should be carried out separately for monolingual Germans and bi-/multilinguals.

Exclusion of bi-/multilinguals indeed changed results radically. Among monolingual German children, those with frequent otitis media were classified more often as needing additional medical assistance in acquiring German (17/68 (25.0%) vs. 94/670 (14.0%); $\chi^2_{(1)} = 5.8$, $p = .016$) and, therefore, they failed marginally significantly more often in KiSS.2 (20/68 (29.5%) vs. 132/670 (19.7%); $\chi^2_{(1)} = 3.6$, $p = .059$). According to parents, such children also had more often language delay (first words after the second birthday; $\chi^2_{(1)} = 4.0$, $p = .046$), apart from other unfavourable medical factors mentioned above. Children with and without frequent otitis media did not differ in age (in months), that is, were comparable. Thus, monolingual German children with frequent otitis media had somewhat weaker German language skills than children without frequent otitis media.

In the subgroup of bi-/multilinguals, the strange finding that frequent otitis media is associated with better German language competence did not disappear, in spite of low sample sizes. Children with frequent otitis media were less often classified as needing educational assistance in acquiring German (22/40 (55.0%) vs. 612/848 (72.2%); $\chi^2_{(1)} = 5.5$, $p = .019$) and they also failed less often in KiSS.2 (22/40 (55.0%) vs. 619/848 (73.0%); $\chi^2_{(1)} = 6.2$, $p = .013$). This can be explained by the fact that in their families German was spoken more often than in the families of children without frequent otitis media (lbl = 6.9, $p = .009$). Again, there was no difference in age between children with and without frequent otitis media. It can be concluded that not otitis media, but the use of German at home resulted in better German language skills.

Results on Dyslexia in the Family

In dichotomized KiSS.2 results, there were no differences between children with and without cases of dyslexia in the family, but Mann-Whitney *U*-test identified minimal differences in KiSS.2 subtests. Whereas children with and without dyslexia in the family scored approximately at the same level in the German articulation, grammar, repetition of non-words and sentences as well as in the KiSS.2 total score, children with dyslexia in the family outperformed children without dyslexia in the family in the speech comprehension (5.4 ±2.6 vs. 5.0 ±2.8; $Z = -2.0$, $p = .048$) and vocabulary (15.5 ±6.3 vs. 12.9 ±7.3; $Z = -2.3$, $p = .022$). Again, this finding can be explained by a higher percentage of monolingual Germans among children with dyslexia in the family, compared to children without dyslexia in the family (115/732 (15.7%) vs. 65/884 (7.4%); $\chi^2_{(1)} = 28.3$, $p < .001$). Thus, children with dyslexia in the family spoke better German not because of dyslexia but because they were more often Germans.

Exclusion of bi-/multilinguals shows that, as expected, monolingual German children with cases of dyslexia in the family needed more often educational (33/115 (28.7%) vs. 85/617 (13.8%); $\chi^2_{(1)} = 16.0$, $p < .001$) and medical support (25/115 (21.7%) vs. 84/617 (13.6%); $\chi^2_{(1)} = 5.0$, $p = .025$) in the German language acquisition, and therefore, they failed more often in KiSS.2 (38/115 (33.0%) vs. 111/617 (18.0%); $\chi^2_{(1)} = 13.6$, $p < .001$). They showed weaker results in all subtests of KiSS.2 except speech comprehension and repetition of non-words (ps < .05). However, even in the repetition of non-words the difference was marginally significant with $p = .068$ (and in the speech comprehension such differences only seldom become statistically significant due to a low number of items: only three). On the contrary, results of bi-/multilinguals were not statistically significant (Chi-Square and Mann-Whitney *U*-tests) because such variables as age of the German language acquisition are more important for them than the familial predisposition for language impairments and disorders.

Results on Children Born in the First vs. Second Half of the Year

According to KiSS.2, in S3 children born in January–July needed less often additional medical assistance in acquiring German than those born in August–December (164/907 (18.1%) vs. 205/874 (23.5%); $\chi^2_{(1)} = 7.8$, $p = .005$). Also, they needed marginally significantly less often educational assistance in acquiring German (368/907 (40.6%) vs. 392/874 (44.9%), $\chi^2_{(1)} = 3.3$, $p = .068$) and failed marginally significantly less often in KiSS.2 (383/907 (42.2%) vs. 405/874 (46.3%), $\chi^2_{(1)} = 3.1$, $p = .081$). Again, this finding is explained by a higher percentage of monolingual Germans in the subgroup of children born in January-July: 1,273/2,577 (49.4%) vs. 766/1,739 (44.0%); $\chi^2_{(1)} = 11.9$, $p = .001$.

In the subgroup of monolingual Germans, differences between children born in January-July and August-December did not exist. However, in the subgroup of bi-/multilinguals, children born in January-July were still less often classified as needing additional medical assistance in acquiring German than children born in August-December: 95/449 (21.2%) vs. 132/481 (27.4%); $\chi^2_{(1)} = 5.0$, $p = .026$. There was only one factor that might have contributed to this difference: mothers of children born in August-December had a lower educational level: lbl = 4.4, $p = .036$. Thus, it would be possible to explain the difference in KiSS.2 by some variables related to the educational level (e.g., more neglect of children born in August-December). It is also noteworthy that children born in the first half of the year were more often tested in the years 2008-2012 (534/1,304 (41.0%) vs. 357/973 (36.7%)), when German language skills generally were better, and children born in the second half of the year were more often tested in the years 2017-2019 ($\chi^2_{(1)} = 4.3$, $p = .039$).

One can reverse the results if another sample is used. In S4, children born in the second half of the year showed better German language competence than children born in the first half of the year. They were less often classified in KiSS.2 as needing educational assistance in acquiring German (610/3,307 (18.4%) vs. 599/2,837 (21.1%); $\chi^2_{(1)} = 6.9$, $p = .009$) and also they failed less often in KiSS.2 (870/3,307 (26.3%) vs. 826/2,837 (29.1%); $\chi^2_{(1)} = 6.0$, $p = .014$). Here, percentages of monolinguals and bi-/multilinguals in the first and second halves of the year did not differ. Unfortunately, this sample contains almost no other sociodemographic or medical data. Therefore, the confounding variable(s) remain(s) unknown. Also, it should be emphasized that samples S3 and S4 used different versions of KiSS (KiSS.1 and KiSS.2), which reduces their comparability.

Results on Stuttering

Stutterers (children stuttering often or always according to the questionnaire for parents) made out 4.8% of the sample ($n = 36$), which is normal for children of the pre-school age, but represents a problem for statistical calculations because a very small subgroup should be compared with a large one.

Correlations between stuttering (ordinal: “never – seldom – sometimes – often – always”) and total scores of correct answers in language tests show that stutterers slightly lagged behind in the German language acquisition (Table 2).

Table 2
Associations Between Stuttering and Total Scores of Language Tests: Spearman’s Correlations

	S-ENS: repetition non-words	S-ENS: repetition sentences	S-ENS: filling gaps in words	S-ENS: articulation	ETS: speech comprehension	ETS: grammar	AWST-R: vocabulary
ρ	-.092*	-.111**	-.057	-.117**	-.126**	-.106**	-.155***
n	581	580	581	579	696	678	684

Note. *** $p < .001$, ** $p < .01$, * $p < .05$

The first correlation in Table 2 (-.092) should be dismissed as too low: according to Cohen (1988), correlations < .1 should not be reported. The second correlation (-.111) would disappear if we apply the Bonferroni adjustment of p-values ($p = .05/7 = .007$) and would not if we apply the Bonferroni-Holm adjustment or do not apply adjustments at all. All other correlations can

be considered weak, but statistically significant even after the adjustment of *p*-values. They show that there was indeed a minimal, but statistically significant association between stuttering and poor German language skills. All *p*-values reported in Table 2 result from two-sided calculations. If we formulate a directional hypothesis ("there must be a negative association between stuttering and German language skills: more stuttering is associated with worse language skills"), then we can use one-sided calculations that deliver more statistically significant results. In case of the repetition of sentences, the *p*-value would go down from .007 to .004 and, thus, it would still remain significant even after the application of the Bonferroni adjustment. Alternatively, one can use Pearson's correlations instead of Spearman's. Pearson's correlations are less conservative and deliver more often significant *p*-values. In case of the repetition of sentences, the *p*-value would go down from .007 to .003 and, again, it can be "saved" from the Bonferroni adjustment. On the other hand, inclusion of more correlations (e.g., total score of S-ENS) would automatically reduce the number of statistically significant results, if we still want to apply the Bonferroni adjustment (.05/8 = .00625). Thus, inclusion of just one more correlation would make us lose a statistically significant result on grammar. Also, different dichotomizations of ordinal variables can change results dramatically. If we dichotomize the variable "stuttering" as "never-seldom vs. sometimes-always" (instead of "never-sometimes vs. often-always"), then the point-biserial correlation with the S-ENS subtest "articulation" is no more significant and too low to be reported (-.082).

Also, minimal differences in sample sizes can change a lot in case of such a seldom phenomenon as stuttering. If an unselected sample is tested, then very few stutterers need to be compared with a lot of non-stutterers. Every new stuttering child can change the results. Thus, if we take into account the date of language tests, then the first 17-21 recruited stutterers showed in Pearson's correlations that there is a significant (negative) association between stuttering and articulation, but after that the association got lost (children 22+). The association between stuttering and repetition of sentences became statistically significant in the last moment, after the last stuttering child was tested. On the contrary, association between stuttering and filling gaps in words got lost after the addition of the last child. Sample sizes of 200, 370, and 580 (all children taken together) would have resulted in the highest number of statistically significant Pearson's correlations. Other sample sizes would deliver lower numbers of significant results. With 100 children, no results would be statistically significant. Thus, between zero and seven results can be "made" statistically significant in univariate tests depending on the sample size.

Instead of correlations, Mann-Whitney *U*-tests can be used to analyze the link between stuttering and language skills. For this purpose, stuttering should be dichotomized (e.g., "never-seldom vs. sometimes-always"). Here, results can change depending on (a) one-sided or two-sided calculations, (b) exact or approximate (asymptotic) calculations of the *p*-value, (c) dichotomization, (d) imputation, (e) application of *p*-value adjustments etc. For those researchers who do not want to rely on *p*-values only, various effect sizes were developed. Here, probability of superiority index was used ($\hat{p}$, Grissom & Kim, 2012). Values close to .5 can be considered a low effect size, those close to .0 high. In Table 3, results show that different calculation methods did not change much in the results. Significance values were, as always, higher in one-sided calculations but it did not make results qualitatively different. Also, we were lucky enough to receive in Mann-Whitney *U*-test results comparable to those in correlations in Table 2. Unfortunately, we cannot completely rely on the effect size because, for instance, a considerable effect size of .29 in the AWST-R vocabulary would not change even if we compare just five stutterers with 700 non-stutterers.

Table 3
Differences Between Children Stuttering Never-Seldom vs. Sometimes-Always in Total Scores of Language Tests: Mann-Whitney U-tests

	S-ENS: repetition non-words	S-ENS: repetition sentences	S-ENS: filling gaps in words	S-ENS: articulation	ETS: speech comprehension	ETS: grammar	AWST-R: vocabulary
Two-sided (z)	-1.27	-2.39*	-1.05	-2.08*	-2.72**	-3.55***	-2.61**
MC two-sided (z)	-1.27	-2.39*	-1.05	-2.08*	-2.72**	-3.55***	-2.61**
MC one-sided (z)	-1.27	-2.39**	-1.05	-2.08*	-2.72**	-3.55***	-2.61***
$\hat{p}$	.43	.37	.44	.40	.36	.31	.29
n	581	580	581	579	696	678	684

Note. *** $p < .001$, ** $p < .01$, * $p < .05$, MC = Monte Carlo (exact calculation of p-values)

It is important to know that various statistical methods demand various minimal sample sizes (that can be calculated in special programs). If this requirement was not fulfilled, results of statistical calculations become unreliable. In our case, it does not matter how we dichotomize stuttering: "never-seldom vs. sometimes-always" or "never-sometimes vs. often-always" because the sample size in the subgroup of stutterers always remains too low for Mann-Whitney U-tests and, therefore, our results remain questionable. Mann-Whitney U-test requires at least 67 participants in each subgroup (two-sided calculations, effect size d .05, power .80, α error probability .05). Ideally, we would need 110 participants in each group (power .95). We cannot calculate sample sizes for a one-sided test because, as was shown above, sometimes stutterers outperform non-stutterers in language tests (one-sided tests demand less participants than two-sided tests).

Table 4 summarizes different univariate statistical methods that can be used for the analysis of associations between stuttering and language skills. In Table 4, the common denominator of all calculations is that the association with grammar skills is always significant (all correlations are significant and above .1, that is, can be reported). Results for all other linguistic domains vary depending on the (subjective) choice of the statistical method. Again, some results can be "made" statistically significant by means of, for instance, imputation. Thus, the T-test on the repetition of sentences did not deliver a significant result in Table 4, but if we use imputation with mean values, then it would become significant: $T = 2.01$, $p = .045$.

Table 4
Differences Between Children Stuttering Never-Seldom vs. Sometimes-Always in Several Language Tests: Various Univariate Statistical Tests

	S-ENS: repetition non-words	S-ENS: repetition sentences	S-ENS: filling gaps in words	S-ENS: articulation	ETS: speech comprehension	ETS: grammar	AWST-R: vocabulary
T-test (*t*)	0.86	1.60	1.32	1.45	3.07**	3.38**	4.74***
U-test (*z*)	-1.27	-2.39*	-1.05	-2.08*	-2.72**	-3.55***	-2.61**
ρ correlation	-.092*	-.111**	-.057	-.117**	-.126**	-.106**	-.155***
r correlation	-.090*	-.124**	-.077	-.120**	-.141***	-.150***	-.162***
φ correlation	.032	.123**	.044	.086*	.056	.126**	.094*
Linear-by-linear	0.60	8.76**	1.11	4.23*	2.16	10.67**	5.95*

Note. *** $p < .001$, ** $p < .01$, * $p < .05$

It should be noted that univariate methods deliver more uniform results than multivariate ones. Thus, with multivariate methods virtually any result is possible. In the current study, only univariate methods were applied.

Discussion

It was shown that if confounding variables are not taken into account, any most absurd result is possible in statistical calculations. Here, this confounding variable was the use of German at home. Children with frequent otitis media, cases of dyslexia in the family, and born in the first half of the year spoke better German simply because they were more often monolingual Germans or used more often German at home (if bi-/multilingual). Not always such confounding variables can be considered a matter of chance. In case of otitis media and dyslexia, one can assume that monolingual Germans as well as integrated immigrants (those who often use German at home) have an easier access to the German healthcare system (cf. Holz, 2022; Lechner & Mielck, 1998; Maier et al., 2015). Therefore, their medical issues are diagnosed more often and quickly (cf. Scharff Rethfeldt, 2019). It is hardly imaginable that bi-/multilinguals have less often otitis media and dyslexia but their medical issues remain longer undiagnosed and, thus, without treatment. This results in worse German language skills. In the example with the first vs. second halves of the year, there was another confounding variable, namely better German language skills of children tested in 2008-2012, compared to those in 2017-2019. Language assessment studies have their active phases that last for several months. In 2008-2012 such an active phase happened to be in the first half of the year, which resulted in inclusion of many monolingual Germans and bi-/multilinguals with comparatively good German language skills in the sample. In 2017-2019, more children were tested in the second half of the year. As was mentioned above, German language skills were much better in 2008-2012 than in 2017-2019. This is valid both for monolingual Germans (Zaretsky et al., 2022) and bi-/multilinguals (Zaretsky et al., 2020).

Example of stuttering shows that even studies with large sample sizes do not guarantee correct results because (a) a large sample size does not mean that the subgroups to be compared are also large (the subgroup of stutterers was very small), (b) the choice of statistical methods can influence results considerably. Especially the use of the Bonferroni adjustment of *p*-values is questionable and has been often criticized as too rigorous (Narum, 2006; Perneger, 1998), but is widely applied in the research. Even in case of relatively simple univariate methods one can present the link between stuttering and language skills as non-existent to considerable. Although language tests used in this study are well-known, validated, and standardized, there is still enough space for manipulations or errors. Among other things, imputation by mean values can make one or two results statistically significant (those that were on the verge of statistical significance before imputation). Imputation also influences some other statistical results (e.g., makes standard deviations smaller). One-sided calculations and parametrical statistical methods automatically make results more statistically significant. Effect sizes can be helpful in the recognition of "unreliable" *p*-values but sometimes do not "notice" other problems such as incomparable sample sizes.

Taking into account that in most studies sample sizes are too low and various adjustments of *p*-values are often applied without any clear motivation, many statistically significant results get lost. These lost results often remain unpublished, that is, invisible for the reader because scientific journals still prefer to accept manuscripts with clear, statistically significant findings. This leads to the so-called *p*-hacking, that is, various manipulations of subsamples, inclusion and exclusion criteria, or parameters of statistical methods with the aim of getting statistically significant results that can be published.

Conclusion

There is a considerable variability in statistical results depending on the choice of statistical methods, sample size, imputation, and many other factors. Also, confounding variables can change results of statistical calculations dramatically.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The authors declare that no AI or AI-assisted technologies have been used to generate, refine, or correct the content in the manuscript. The ideas, design, procedures, findings, analyses, and discussion are originally written and derived from careful and systematic conduct of the research.

References

Angermaier, M. J. W. (2007). *ETS 4-8. Entwicklungstest Sprache für Kinder von 4 bis 8 Jahren* [ETS 4-8. Developmental test for children aged between 4 and 8 years]. Harcourt Test Services GmbH.

Caglar-Ryeng, O., Eklund, K., & Nergard-Nilssen, T. (2019). Lexical and grammatical development in children at family risk of dyslexia from early childhood to school entry: A cross-lagged analysis. *Journal of Child Language, 46*(6), 1102–1126. https://doi.org/10.1017/S0305000919000333

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* 2nd ed. Erlbaum.

Döpfner, M., Dietmair, I., Mersmann, H., Simon, K., & Trost-Brinkhues, G. (2005). *S-ENS. Screening des Entwicklungsstandes bei Einschulungsuntersuchungen* [S-ENS. Screening of the developmental status for school enrolment examinations]. Hogrefe.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications*. 2nd ed. Routledge.

Holler-Zittlau, I., Euler, H. A., & Neumann, K. (2011). Kindersprachscreening (KiSS) – das hessische Verfahren zur Sprachstandserfassung [Language Screening for Children (KiSS) – a Hessian language screening tool]. *Sprachheilarbeit, 5*(6), 263–268.

Holz, M. (2022). Health inequalities in Germany: Differences in the 'Healthy migrant effect' of European, non-European and internal migrants. *Journal of Ethnic and Migration Studies*, *48*(11), 2620–2641. https://doi.org/10.1080/1369183X.2021.1901675

Kiese-Himmel, C. (2005). *AWST-R. Aktiver Wortschatztest für 3- bis 5-jährige Kinder – Revision* [AWST-R. Test of active vocabulary for children aged between 3 and 5 years – revision]. Hogrefe.

Lechner, I., & Mielck, A. (1998). Die Verkleinerung des "Healthy-Migrant-Effects": Entwicklung der Morbidität von ausländischen und deutschen Befragten im sozioökonomischen Panel 1984-1992 [Decrease in the "healthy migrant effect": Trends in the morbidity of foreign and German participants in the 1984-1992 Socioeconomic Panel]. *Das Gesundheitswesen, 60*(12), 715–720.

Maier, I., Kriston, L., Härter, M., Hölzel, L. P., & Bermejo, I. (2015). Psychometrische Überprüfung eines Fragebogens zur Erfassung der Barrieren der Inanspruchnahme von Gesundheitsleistungen durch Personen mit Migrationshintergrund [Psychometric testing of a new scale assessing the reasons for non-utilisation of health care services by people with migration background]. *Gesundheitswesen, 77*(10), 749–756. https://doi.org/10.1055/s-0034-1395641

Narum, S. R. (2006). Beyond Bonferroni: Less conservative analyses for conservation genetics. *Conservation Genetics, 7*, 783–787. https://doi.org/10.1007/s10592-005-9056-y

Nittrouer, S., & Lowenstein, J. H. (2024). Early otitis media puts children at risk for later auditory and language deficits. *International Journal of Pediatric Otorhinolaryngology, 176*, 111801. https://doi.org/10.1016/j.ijporl.2023.111801

Ntourou, K., Conture, E. G., & Lipsey, M. W. (2013). Language abilities of children who stutter: A meta-analytical review. *American Journal of Speech-Language Pathology, 20*(3), 163–179. https://doi.org/10.1044/1058-0360(2011/09-0102)

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal, 316*, 1236. https://doi.org/10.1136/bmj.316.7139.1236

Reilly, S., Onslow, M., Packman, A., Cini, E., Conway, L., Ukoumunne, O. C., Bavin, E. L., Prior, M., Eadie, P., Block, S., & Wake, M. (2013). Natural history of stuttering to 4 years of age: A prospective community-based study. *Pediatrics, 132*(3), 460–467. https://doi.org/10.1542/peds.2012-3067

Scharff Rethfeldt, W. (2019). Speech and language therapy services for multilingual children with migration background: A cross-sectional survey in Germany. *Folia Phoniatrica et Logopaedica, 71*(2-3), 116–126. https://doi.org/10.1159/000495565

Zaretsky, E., Lange, B. P., Euler, H. A., Robinson, F., & Neumann, K. (2017). Pre-schoolers who stutter score lower in verbal skills than their non-stuttering peers. *The Buckingham Journal of Language and Linguistics, 10*, 96–115.

Zaretsky, E., van Minnen, S., Lange, B. P., & Hey, C. (2020). Sprachkompetenzen vierjähriger Kinder mit Migrationshintergrund in Hessen: eine Bestandsaufnahme [Language competence of children with immigrant background in Hesse: A survey]. *Praxis Sprache, 65*(2), 90–97.

Zaretsky, E., van Minnen, S., Lange, B. P., & Hey, C. (2022). Sprachstand vierjähriger monolingual deutscher Kinder: eine Querschnittsanalyse [Language competence of four-year-old monolingual German children: A cross-sectional analysis]. *Kindheit und Entwicklung, 31*(1), 52–59. https://doi.org/10.1026/0942-5403/a000363