Comparison of self-administered versus read-aloud questionnaires for psychological measurement in students with low intellectual functioning: use of frequentist and Bayesian approaches

Claudia P. Pérez-Salas[1], Victoria Parra[2], Alonso Ortega[3], Fabiola Sáez-Delgado[4], Pamela Ramírez-Peña[5] and Isidora Zañartu[1]

[1]Social Science Faculty, Psychology Department, Universidad de Concepción, Concepción, Chile; [2]Independent Researcher, Santiago, Chile; [3]Phonoaudiology Department, Universidad de Valparaíso, Valparaíso; [4]Education Faculty, Fundamentals of Pedagogy Department, Universidad Católica de la Santísima Concepción, Concepción, Chile; [5]Humanities Faculty, Spanish Department, Universidad de Concepción, Concepción, Chile

Background: Students with low intellectual functioning (LIF) often experience barriers to participating in social research due to the literacy demands of the survey's typical self-administered format. Although evidence for the validity of the read-aloud format for educational testing abounds, few studies have analyzed the impact of application formats on attitudes or opinion questionnaires for LIF students.
Aim: To analyze the effect of self-administered vs read-aloud formats on LIF and typical development (TD) students using four psychological questionnaires for school contexts (Student Engagement Instrument, Multidimensional School Engagement Scale [MSES], Brief Multidimensional Student Life Satisfaction Scale, and School Participation Scale).
Method: A mixed factorial (2x2) design was used. Thirty-two students participated (14 to 19 years old; $M = 15.39$; $SD = 1.27$): 17 with LIF and 15 with TD.
Results: Reliability indices between formats for LIF students in most questionnaire subscales were found to be adequate and equivalent. All instrument subscales had appropriate intra-subject correlations between formats, indicating that LIF students had similar scores in both. Only the MSES showed a format effect, where LIF students reported fewer disengagement behaviors in the read-aloud format. Frequentist and Bayesian statistics were conducted looking for convergences due to the small sample size.
Conclusion: We discuss the case-related appropriateness of each application format and propose a new criterion to choose between them to guarantee the inclusion of LIF students in psychological research.

Keywords: Self-administered; read-aloud; assessment format; low intellectual functioning; questionnaires; bayesian statistic

Introduction

A widely used standard procedure for psychological measurement is the self-administered questionnaire. However, it has been difficult to apply this method to people with low intellectual functioning[1] (LIF; Finlay and Lyons 2001, Goegan *et al.* 2018), and there is scant empirical evidence about the best and most reliable way to implement it in this population (Kooijmans *et al.* 2022, Shogren *et al.* 2021).

Some scholars argue that self-administered questionnaires can be inappropriate for assessing people with LIF, because of difficulties in a) understanding the meaning of the questions; b) the cognitive processing to respond (e.g. recalling or comparing information); and c) the ability to express a clear answer (e.g. choosing a multiple-choice answer) (Emerson *et al.* 2013, Finlay and Lyons 2001).

However, Redline *et al.* (2005) stated that some difficulties can also arise when assessing the general population with some self-administered questionnaires, specifically those with branching instructions, because of participants' tendency to ignore, misread, or not appropriately follow these instructions, leading to item omissions or erroneous completion. Using an

Correspondence to: Claudia P. Pérez-Salas, Facultad de Ciencias Sociales, Departamento de Psicología, Universidad de Concepción, Barrio Universitario s/n, Concepción 4070112, Chile. Email: cperezs@udec.cl

DOI 10.1080/20473869.2023.2243414

experimental design, Redline *et al.* (2005) assessed 1,266 undergraduate typical development (TD) university students and concluded that five of the seven question-complexity characteristics (i.e. high number of answer categories, all categories' branches, write-in responses, location at the bottom of the page, and high distance between the answer box and branching instruction) increased errors of commission, that is when the respondent answered a question that they were not instructed to answer. Contrary to the prediction, only two characteristics of question complexity (i.e. write-in response and location at the bottom of the page) increased errors of omission (i.e. respondents leaving questions blank). These findings support the idea that both, visual and verbal complexity of information on a self-administered questionnaire, may affect either what respondents read, their order of reading or their comprehension of the information.

The reading-aloud format involves either an in-person interviewer or a recording of the questionnaire being read-aloud and the subject's responses (Buzick 2019, Buzick and Stone 2014). The latter is a frequent accommodation for the evaluation of people with learning disabilities and/or LIF (Sireci *et al.* 2018, Thurlow *et al.* 2012) because it reduces the cognitive burden by eliminating the requirement for the respondent to be literate (Bowling 2005) or have good reading comprehension (Gresch *et al.* 2016). The characteristics of this method suggest that it could be beneficial for all students (Li 2014) as it reduces the reading requirements for respondents (Wood *et al.* 2018).

Reading aloud is widely used in large-scale educational evaluations of people with literacy limitations to remove barriers when demonstrating their knowledge (Buzick 2019, Buzick and Stone 2014). The data scores of self-administered versus read-aloud questionnaires have been primarily evaluated in specific content areas (Rogers *et al.* 2019), with most studies focusing on mathematics and reading comprehension of elementary and secondary school students (e.g. Giusto and Ehri 2019, Spiel *et al.* 2016). These studies state that reading aloud could help measure a given topic more effectively in students with learning disabilities, reducing measurement error by eliminating the variance linked to decoding in the assessment of math, or reading comprehension, promoting fairer assessments (Andreou *et al.* 2019, Buzick 2019).

The interaction hypothesis was used to justify the validity of the accommodations in the performance tests. This hypothesis states that accommodation is useful only for increasing the scores of people who require accommodation, but not for those who do not need it (Sireci *et al.* 2005, Zuriff 2000). A psychometric explanation of the latter is that some aspects of standardized test administration introduce irrelevant variance for the evaluated construct when assessing people with disabilities, hence, their elimination would lead to a more valid result. Regarding non-performance tests, data quality (e.g. missing data, correct understanding of filter questions), and concurrent validity have been analyzed to understand the equivalence between formats. However, the literature is scarce (see Chang and Krosnick 2010, Gresch *et al.* 2016).

The meta-analyses conducted by Buzick and Stone (2014) and Li (2014) on the effects of read-aloud accommodation for all students concluded that the read-aloud accommodation failed to meet the interaction hypothesis criteria for accommodation validity, and supported a softer version of it named as 'differential boost hypothesis' (Fuchs and Fuchs 2001). This hypothesis states that accommodation may be appropriate even if TD students would benefit from it, as long as students with disabilities who need accommodation receive a significantly greater score increase. However, more research is required to establish the validity of inferences based on the read-aloud administration scores and the comparability of scores with and without accommodations (Li 2014, Buzick and Stone 2014).

There is growing evidence that accommodations can remove variance related to test features, and modifications to testing conditions are common in schools to measure the performance of students with learning disabilities, specifically, in math and reading comprehension. Despite the latter, the use of accommodations in social research has received little attention (Goegan *et al.* 2018). Hence, few studies have evaluated the possible effect of testing accommodations on the data quality with respects to constructs such as attitudes, perceptions and/or opinions.

An exception is the study of Gresch *et al.* (2016) that implemented an experimental intra-subject design to assess the effect of a read-aloud format versus the self-administered format on the quality of responses to a short version of the German National Educational Panel Study (NEPS) Questionnaire. The participants were 664 12-year-old secondary students from a Hauptschule in Germany (which usually comprises students with low achievement and poor readers). In this study, fifth grade students were randomly assigned to either self-administered or read-aloud treatment conditions. The NEPS questionnaire includes items on the students' backgrounds, grades, reading habits, and self-esteem. The first group completed the questionnaires individually, whereas in the second the administrator read aloud each item using a standardized script. The day after, experimental conditions were reversed. Every five minutes, all students in both conditions were asked to circle the number of items they were responding to, in order to measure the speed of questionnaire completion. After 15 min, all students were requested to stop filling out the questionnaires. Results indicate higher data quality in the read-aloud condition (i.e. fewer non-

response items, a better understanding of filter questions), which was independent of the students' reading speed and migrant status. The authors concluded that this form of response could benefit students at academic or social risk.

Chang and Krosnick (2010) used political data to compare the effect of different assessment formats. Although this research does not include people with LIF but TD undergraduate university students, it is important because it included a measure of the cognitive skills of participants (i.e. verbal and mathscores on SAT or ACT) to evaluate whether the format effect interacts with them. Using an experimental design, Chang and Krosnick (2010) compared the effect on the data quality of the read-aloud questionnaire and a self-administered questionnaire. They assessed political opinions in a sample of 332 students (174 males and 158 females) attending to an introductory psychology course. The authors were also interested in the presence of a moderating effect of cognitive skills on the format, as reading aloud could impose a greater challenge for people with limited cognitive skills by adding cognitive burden; subjects would have to hold a question and response in their working memory while searching for the answer in long-term memory and generating a judgment. However, visual presentation in self-administered format could reduce this burden and hence help people with limited cognitive skills (Chang and Krosnick 2010).

Chang and Krosnick's (2010) results showed higher concurrent validity, and less satisficing and social desirability response bias in the self-administered format. These format differences were most pronounced among the respondents with limited cognitive skills. Hence, as hypothesized, participants with higher cognitive skills could manage the two formats well, but those with limited cognitive skills were more challenged by the oral presentation. The lower social desirability in the self-administered condition is explained by the authors because of the absence of a human interviewer, making examinees more willing to provide honest but not socially admirable answers (Chang and Krosnick 2010). The same effect of self-administered questionnaires (i.e. mail, paper, pencil, or computerized) on social desirability compared to the read-aloud format (i.e. telephone or in-person interviewer) has been reported in sensitive topics, such as drinking behavior (De Leeuw 1992) and pain (Lozano *et al.* 2016), in the TD population.

The studies by Gresch *et al.* (2016) and Chang and Krosnick (2010) are important to understand that different formats could lead to different data quality, specifically when questionnaires include filter (i.e. complex) questions and have time limits that can increase the risk of missing or erroneous data. However, these studies showed some limitations. In the case of Gresch *et al.* (2016), when measuring data quality, it only focused on omitted answers and a correct understanding of filter questions but did not explore the reliability of the answers in each format. On the other hand, Chang and Krosnick (2010) assessed concurrent validity, but they did not assess the reliability indices nor the convergent validity of formats.

From the literature review, we can conclude that most research on the format effect in students with LIF has focused on math and reading tests, and that there is limited evidence regarding other constructs, such as attitudes or opinion questionnaires, even in TD population. The scant social research compares questionnaires' raw scores in self-administered versus read-aloud formats or analyzes data quality (i.e. omissions or errors) in both conditions, but disregards psychometric properties, such as reliability and convergent validity between formats. Finally, few studies have compared the format effect between students with LIF and TD students, making it difficult to understand whether the format interacts with the student condition.

Thus, the present study attempts to fill this gap by analyzing the effect of self-administered and read-aloud formats on the scores presented by LIF and TD students on four instruments that measure attitudes and feelings in the school context (Student Engagement Instrument, Multidimensional School Engagement Scale (MSES), Brief Multidimensional Student Life Satisfaction Scale, and School Participation Scale). Subscale raw scores of each instrument were compared within and between groups, along with reliability and convergent indices between the formats.

Methods

A mixed quasi-experimental 2×2 factorial design was used. The intra-subject variable was the application format (i.e. self-administered versus read-aloud), and the inter-group variable was the student's condition (i.e. LIF versus TD). This design is a quasi-experiment because students cannot be randomly assigned to the LIF or TD conditions, since they are intrinsic variables, but there was an intentional manipulation of the intra-subject variable.

Participants

A group of 32 male and female Chilean students between 14 and 19 years of age ($M = 15.39$; $SD = 1.27$), were divided into two subpopulations: LIF ($n = 17$; 71% women) and TD students ($n = 15$; 53% women). Among the former, 10 had a diagnosis of borderline intellectual functioning, and seven had a diagnosis of mild intellectual disability. All of them had also deficits in their adaptive behavior. The diagnosis was provided by the school inclusion program, which is guided by an Educational Ministry of Chile protocol which is based on the multidimensional model of intellectual disability

of the World Health Organization, and the ICD-10 classification. This protocol states that diagnosis must be made by a specialized psychologist and comprises three criteria: intellectual functioning limitations (assessed exclusively with WISC test); adaptive functioning limitations (assessed with a standardized test) and onset before the age of 18. For the diagnosis of mild intellectual disability, an IQ ranging from 50 to 69 is required, as well as a score below the mean in the adaptive behavior assessment by 2 standard deviations. For the diagnosis of borderline intellectual functioning, an IQ ranging from 70 to 79 is required, as well as a score below the mean in the adaptive behavior assessment (Mineduc 2009).

The inclusion criteria for the LIF student sample were: (a) to be enrolled in the 9th or 10th grade, (b) to be part of the school integration program, and (c) to have a mild intellectual disability or borderline intellectual functioning diagnosis. The inclusion criteria for the TD group were (a) to be enrolled in 9th or 10th grade, and (b) to have no intellectual disabilities or belonging to the school integration program.

Psychiatric disorders (i.e. schizophrenia, bipolar disorder, and autism spectrum disorder) or multiple deficits were part of the exclusion criteria for both subsamples. The school provided information regarding diagnoses to verify compliance with the inclusion criteria.

When analyzing the composition of the sample, no differences were found in the proportion of men and women by group, $\chi^2_{(1)} = 1.012$, $p = .314$, or age between groups ($t_{(21)} = .882$; $p = .388$).

Regarding economic status, 93.3% of the TD and 81.3% of the LIF sample had a family income below 650 USD, which corresponds to a low socioeconomic status in Chile.

Instruments and variables

Independent variables

Intellectual functioning. (a) Low intellectual functioning (LIF). It is an umbrella term to refer to students that have an IQ below 85 (Lees-Warley and Rose 2015). This study includes people with mild intellectual disability and borderline functioning, and (b) Typical Development (TD). Describes students that do not have low intellectual functioning.

Format of application. (a) Self-administered questionnaire, and (b) reading-aloud assessment.

Dependent variables

The Student Engagement Instrument (SEI). It was developed by Appleton and Christenson in 2004 (Appleton *et al.* 2006). It consists of 35 items measuring six subscales of school engagement: (a) teacher–student relationships [TSR] (nine items: 'My teachers are there when I need them'), (b) control and relevance of school work [CRSW] (nine items: 'After finishing my schoolwork I check to see if it is correct'), (c) peer support at school [PSS] (six items: 'Other students at school care about me'), (d) future aspirations and goals [FAG] (five items: 'I plan to continue my education after high school'), (e) family support for learning [FSL] (four items: 'My family/guardian(s) want me to keep trying when things are tough at school'), and (f) intrinsic motivation (inverted) [IM] (two items: 'I'll learn, but only if my family/guardian(s) give me a reward'). Each item is answered on a 4-point Likert scale (strongly disagree, disagree, agree, and strongly agree).

Appleton *et al.* (2006) demonstrated that the best fit consists of six factors. These factors are correlated with expected educational outcomes (e.g. academic achievement, graduation, and retention). This instrument was validated by González et al. (2022) in a sample of Chilean students. Internal consistency is acceptable, with omegas greater than .76 in each of the subscales. A six-dimensional factorial structure was also replicated.

The Multidimensional school engagement Scale (MSES). This instrument, developed by Wang *et al.* (2019), consists of 37 items measuring two factors: engagement and disengagement. The school engagement factor contains 19 items in total along four dimensions: (a) behavioral engagement (four items: 'I try my best at school'), (b) cognitive engagement (five items: 'I work hard during challenges/difficulties at school'), (c) emotional engagement (five items: 'I am happy at school'), and (d) social engagement (five items: 'I enjoy working with peers at school'). The disengagement factor contains 18 items with four dimensions: (a) behavioral disengagement (eight items: 'I get into trouble at school'), (b) cognitive disengagement (two items: 'I don't pay attention in class'), (c) emotional disengagement (four items: 'I feel overwhelmed by my schoolwork'), and (d) social disengagement (four items: 'I do not care about people at my school'). The validation study showed a multidimensional bifactorial structure and provided evidence of measurement invariance, construct validity, and predictive validity (Wang *et al.* 2019). Psychometric properties have been evaluated among Chilean students (Pérez-Salas 2021), showing adequate reliability and construct validity index.

The school Participation Scale (SPS). This instrument was developed by John-Akinola and Nic-Gabhainn (2014) and consists of 25 items that measure four dimensions of school participation: (a) in school decisions and rules [PSDR] (six items: 'Students participate in developing the school rules'), (b) in non-academic extracurricular activities [PNEA] (seven items: 'I

have fun doing these extracurricular activities'), (c) in school events [PSE] such as sports day (six items: 'Students participate in planning school events'), and (d) positive perception of school participation [PPSP] (six items: 'At my school, all students get the opportunity to participate'). Each item is answered on a 5-point Likert scale (never to always), with a higher score indicating higher school participation.

This instrument was validated in Chilean school students by Pérez-Salas et al. (2019), with adequate internal consistency indices ranging from $\Omega = .83$ to $\Omega = .87$. Confirmatory factor analysis confirmed the presence of a general school participation factor along with the four dimensions.

The Brief Multidimensional Student Life Satisfaction Scale (BMSLSS). This five-item instrument was developed by Seligson *et al.* (2003) to measure life satisfaction in different areas of life: family, friends, school, self, and life in general (e.g. 'I would describe my satisfaction with my friendships as … '). The Spanish version of each item is answered on a 10-point Likert scale (from completely dissatisfied to completely happy), with higher scores indicating greater satisfaction (Casas *et al.* 2012). Seligson *et al.* (2003) obtained an alpha coefficient of .75 for the BMSLSS in a sample of teenagers and item-total correlations ranging from .65 to .73.

This instrument was validated in Chilean school students by Alfaro *et al.* (2015). The results indicate a reliability coefficient of $\alpha = .70$, and the CFA replicated the presence of a single factor, which is consistent with similar studies. Following Alfaro *et al.* (2015), we included the additional item (item six) to measure global satisfaction in life in general.

Procedure

The ethical, bioethical and biosecurity committee of the Universidad de Concepción, Chile approved this study (CEBB 1092-2022). Institutional authorization was requested from the municipal education department of Temuco, Chile. Through them, we contacted local schools that had integration projects. Two schools showed interest in participating in this study. We contacted both and were able to meet and continue the project with one of them. After this, eligible participants were determined according to the study inclusion criteria for both samples (i.e. students with LIF or TD).

An invitation to participate was sent to the parents of the eligible participants. After explaining the students' rights and the purpose of the study, we obtained active informed consent from the parents and student-informed assent before the evaluation.

The two groups of students were randomly assigned to two quasi-experimental conditions: 19 were first assigned to the self-administered questionnaire format (LIF = 11; TD = 8) and 13 were assigned the read-aloud questionnaire format (LIF = 6; TD = 7). Approximately five weeks later, the same students were evaluated using the other format.

One of the researchers conducted all the evaluations at the participants' school. None of the participants were familiar with her. In the self-administered condition, the test administrator told the students to complete the questionnaires individually, and in the read-aloud condition, each item was read aloud to students using a standardized script. The application was done collectively in a classroom for the self-administered format and individually in a school office for the read-aloud condition. Students were asked to give an answer to all the questionnaire items. There was no time limit for either condition. The average application time for both formats was approximately 45 min. All participants received a movie ticket for collaboration.

Data analysis

Before conducting the data analyses, compliance with parametric test assumptions was tested: (a) normal distribution with asymmetry and kurtosis, and (b) homogeneity of variances with the Box's M Test and Levene's Test. The possible order effect generated by the counterbalance of the two intra-subject conditions was assessed using a factorial ANOVA test of repeated measures. Then, the percentage of missing data was evaluated by application format and by group, and missing data were replaced with the mean to enable analysis with all available cases (i.e. mean imputation of missing data). Descriptive and reliability statistics were performed for each instrument's subscales for both assessment formats with Cronbach's Alpha and Mc Donald's Omega. Also, the correlation between formats was analyzed and Cronbach's Alpha reliability coefficients obtained by each group on both formats were compared.

Last but not least, to evaluate possible differences between the type of applied format (i.e. self-administered vs. read-aloud) and the groups' interaction (i.e. LIF vs. TD), we performed multivariate analyses of mixed variance (i.e. mixed MANOVA) with three of the instruments and a mixed ANOVA with one of them. We chose a mixed MANOVA for the SEI, the MSES, and the SPS since they are multidimensional instruments, and a mixed ANOVA for the BMSLSS since it is a unifactorial questionnaire.

Simultaneously, we conducted Bayesian analyses to complement the frequentist analyses described above. A benefit of conducting Bayesian analyses is that allows us not only to provide evidence against the null hypothesis, as a frequentist approach does, but also in favor of it (Kass and Raftery 1995, Ortega and Navarrete 2017), since 'falsification of the theory requires that one is

able to quantify the support in favor of the null hypothesis' (Wetzels *et al.* 2009, p. 759). Building on this idea, Bayesian analysis allows us to compare two (or more) competing models (e.g. null vs alternative) in light of the existent data and not only based on theoretical probability distributions, as in the Frequentist approach to hypothesis testing (Dienes 2011). An advantage of a Bayesian approach is the use of Bayes Factors (BF) as an index that quantifies the amount of evidence, both for and against, the null hypothesis has, and can be interpreted as an odds ratio which estimates how many times the data is more likely to fit under the null model (i.e. 0) over the alternative model (i.e. 1), and vice versa. As an example, a $BF_{01} = 18$ indicates that it is 18 times more likely that the data supports the null model rather than the alternative one, which is considered strong evidence (see Lee and Wagenmakers 2014). Finally, an additional advantage of Bayesian approaches is that they apply equally validly to any sample size no matter how small (see Jaynes 2003, Jeffreys 1984, Ortega *et al.* 2012). As stated by Wetzels et al. using indices such as 'Bayes Factors (BF) allows us to estimate effect sizes and results particularly appropriate when small numbers of participants or trials are involved' (2011, p. 296).

Analyses were performed with JASP 0.8.4 software (Team JASP 2018) and comparisons between alpha coefficients were conducted with cocron, a platform-independent R package (Diedenhofen and Musch 2016).

Results

Asymmetry values were lower than |2| and Kurtosis values were less than |3| in all instrument subscales for both application formats, supporting the compliance of the assumption of the normal distribution of the variables. Homoscedasticity of both samples and formats was observed using Box's M and Levene's tests for all instruments: Box's $M = 125.446$; $F(78, 2742.9) = .893$; p= .738; all subscales Levene's tests with $p > .05$ [SEI]; Box's $M = 70.213$; $F(36, 2921.4) = 1.378$; $p = .067$; all subscales Levene's tests with $p > .05$ [SPS]; Box's $M = 8.401$; $F(3, 369179.5) = 2.597$; $p = .051$; all subscales Levene's tests with $p > .05$ [BMSLSS]; Box's $M = 9.558$; $F(10, 4135.4) = .816$; $p = .614$; all subscales Levene's tests with $p > .05$ [MSES].

Data analysis showed that the counterbalance was useful, as there was no main effect attributable to the order of application of formats or interaction effect between the order of application and type of format for any of the instruments (Appendix at https://figshare.com/s/de2b5046649e3aa2fac6) .

Participants with different intellectual functioning (intellectual disability and borderline) did not differ on dependent variables, and no interaction effect was found between intellectual functioning level and format in any of the instrument subscales. Thus, we decided to treat all LIF participants as one group.

Considering the 103 items (i.e. summation of the four instruments' items), the total missing values in the LIF group were less than 2% for both formats. In the TD group, the total missing values were less than 2% in the self-administered format and 8.2% in the read-aloud condition. The total number of people with complete values for all items was higher in the self-administered format in the LIF sample. The TD group had the same number of people with complete values for all items in both formats (see Table 1).

Table 2 shows adequate internal consistency indices for SPS, MSES, and BMSLSS subscales in both format conditions for the LIF group (all of them with $\alpha > .75$). The same can be observed for the TD group (except PSDR subscale, $\alpha = .676$). Regarding SEI subscales, internal consistency indices for LIF students range from $\alpha = .677$ and .837, four out of six with $\alpha > .75$) [read-aloud condition]; and $\alpha = .287$ and $\alpha = .857$, two out of six with $\alpha > .75$) [self-administered condition], whereas internal consistency indices for TD students range from $\alpha = .528$ and $\alpha = .873$, five out of six with $\alpha > .75$) [read-aloud condition] and $\alpha = .333$ and $\alpha = .812$, three out of six with $\alpha > .75$) [self-administered condition].

When comparing Alpha coefficients, there were no statistically significant differences between formats in the LIF group for any of the instruments, except for the engagement subscale of MSES. Although both formats presented good reliability indices in this subscale, the read-aloud format was even better ($\chi^2(1) = 9.38$, $p = .002$). For the TD group, there were no statistically significant differences in the reliability of both formats either, except for the peer support at school subscale of SEI. For this subscale, only the self-administered format had an adequate alpha coefficient ($\alpha = .795$) and was significantly better than the alpha coefficient of the read-aloud condition ($\alpha = .528$) ($\chi^2(1) = 4.29$, $p = .038$).

Table 1. Number of complete and missing values by format and group.

	Total missing values per group ($n_{items} = 103 \times n_{group}$)		Total respondents with complete values in their group	
Group	Read-aloud $n_{missing}$ (%)	Self-administered $n_{missing}$ (%)	Read-aloud *n* (%)	Self-administered *n* (%)
LIF ($n = 17$)	27 (1.54%)	16 (0.91%)	7 (41.12%)	10 (58.82%)
TD ($n = 15$)	128 (8.28%)	18 (1.17%)	8 (53.33%)	8 (53.33%)
Total ($N = 32$)	155 (4.70%)	34 (1.03%)	15 (46.88%)	18 (56.25%)

Note. LIF = low intellectual functioning group; TD = typical development group.

Table 2. Descriptive statistics and reliability of instruments according to format application and group.

Instrument[a] (*n* items)	Group (*n*)	Read-aloud M (SD)	Self-admin. M (SD)	*r*. Pearson	BF10[b]	Read-aloud Alpha (Ω)	Self-admin. Alpha (Ω)	Alpha Comparison ($\chi 2(df = 1)$)
SEI								
TSR ($n = 9$)	LIF ($n = 17$)	27.38 (4.63)	27.73 (3.18)	.403	.989	.817 (.830)	.633 (.678)	1.77
	TD ($n = 15$)	30.17 (4.19)	29.62 (3.58)	.823 **	197.67	.859 (.888)	.812 (.839)	.69
PSS ($n = 6$)	LIF ($n = 17$)	18.55 (2.53)	18.32 (3.02)	.734**	52.37	.677 (.702)	.799 (.808)	1.33
	TD ($n = 15$)	19.26 (1.93)	18.78 (3.28)	.802**	113.03	.528 (.621)	.795 (.824)	4.29*
FSL ($n = 4$)	LIF ($n = 17$)	13.13 (2.51)	14.12 (1.58)	.674**	17.39	.759 (.789)	.454 (.537)	2.65
	TD ($n = 15$)	14.56 (1.96)	15.13 (1.30)	.671**	9.76	.873 (.910)	.727 (.778)	2.02
CRSW ($n = 9$)	LIF ($n = 17$)	28.24 (3.52)	28.18 (2.30)	.189	.38	.689 (.745)	.287 (.389)	2.18
	TD ($n = 15$)	29.41 (3.99)	29.62 (3.68)	.766**	48.74	.746 (.814)	.619 (.692)	1.07
FAG ($n = 5$)	LIF ($n = 17$)	17.66 (2.83)	17.92 (2.59)	.360	.764	.837 (.850)	.857 (.873)	.05
	TD ($n = 15$)	18.04 (1.98)	18.11 (1.71)	.717**	19.80	.712 (.843)	.333 (.545)	3.04
IM ($n = 2$)	LIF ($n = 17$)	4.18 (1.70)	4.07 (1.44)	.464	1.53	.800 (.800)	.622 (.632)	.59
	TD ($n = 15$)	3.52 (1.81)	4.40 (2.13)	.370	.74	.762 (.772)	.658 (.660)	.15
SPS								
PSDR ($n = 6$)	LIF ($n = 17$)	21.50 (5.35)	21.35 (4.76)	.715**	35.77	.856 (.868)	.783 (.799)	.94
	TD ($n = 15$)	23.81 (3.74)	25.24 (3.78)	.499	1.64	.676 (–)	.805 (.835)	.82
PSE ($n = 6$)	LIF ($n = 17$)	23.53 (4.73)	22.81 (4.79)	.893**	11903.00	.854 (.890)	.815 (.855)	.76
	TD ($n = 15$)	25.34 (4.22)	26.54 (3.78)	.607*	4.43	.822 (.838)	.812 (.850)	.01
PNEA ($n = 7$)	LIF ($n = 17$)	23.56 (6.58)	24.41 (6.04)	.774**	135.19	.898 (.904)	.837 (.852)	1.57
	TD ($n = 15$)	22.54 (8.22)	23.89 (8.32)	.719**	20.45	.936 (.939)	.909 (.914)	.65
PPSP ($n = 6$)	LIF ($n = 17$)	23.26 (4.81)	23.06 (5.57)	.925**	107283.44	.906 (.908)	.904 (.910)	.01
	TD ($n = 15$)	23.34 (5.27)	24.33 (5.52)	.622*	5.22	.841 (.872)	.843 (.860)	.01
BMSLSS ($n = 6$)								
	LIF ($n = 17$)	49.96 (8.97)	49.88 (9.12)	.400	.97	.750 (.822)	.820 (.836)	.36
	TD ($n = 15$)	48.80 (10.03)	49.13 (10.20)	.891**	2493.53	.847 (.857)	.807 (.846)	.63
MSES								
ENG ($n = 19$)	LIF ($n = 17$)	69.33 (15.55)	71.21 (8.89)	.612**	7.00	.959 (.963)	.838 (.858)	9.38**
	TD ($n = 15$)	74.38 (11.82)	74.80 (9.43)	.593*	3.83	.924 (.936)	.873 (.893)	1.25
DISENG ($n = 18$)	LIF ($n = 17$)	38.37 (10.14)	48.47 (14.72)	.437	1.25	.841 (.864)	.925 (.930)	2.39
	TD ($n = 15$)	36.13 (11.48)	41.06 (14.18)	.797**	99.79	.888 (.905)	.898 (.900)	.07

Note. LIF = low intellectual functioning group; TD = typical development group. [a] SEI: student engagement instrument (subscales: TSR; PSS; FSL; CRSW; FAG; IM); SPS: school participation scale (subscales: PSDR; PSE; PNEA; PPSP); BMSLSS: brief multidimensional student life satisfaction scale; MSES: multidimensional school engagement scale (subscales: Engagement [ENG] and disengagement [DISENG]). [b] Bayesian probability Factor of alternative over null hypothesis.
*p < 0.05; ** p < 0.01.

The correlation between the application formats was direct for all instrument subscales (SEI, SPS, MSES, and BMSLSS) for the LIF and TD groups. For most of the SPS, MSES, and BMSLSS subscales, the correlations' effect sizes ranged from medium to large (Cohen 1988) with most Bayesian Factors indicating 'moderate' or 'strong' evidence in favor of the alternative hypothesis. Correlations between formats were not statistically significant for two subscales in the LIF group (BMSLSS and Disengagement dimension of MSES) and for one subscale in the TD group ('participation in school decisions and rules' dimension of the SEI). In these cases, Bayesian Factors were $BF_{10} = .966$, $BF_{10} = 1.247$ and, $BF_{10} = 1.643$ respectively, all of which can be interpreted as 'anecdotical evidence', and are also consistent with the observed results under the frequentist paradigm.

Regarding SEI scores to both formats, the correlations' effect sizes in the LIF sample ranged from medium to large (Cohen 1988), except for the 'control and relevance of school work' subscale, which was small. Moreover, four of the six subscales did not show statistically significant correlations between formats. In the TD sample, the correlation effect sizes ranged from medium to large. Only one of the six subscales' correlations was not statistically significant. Bayesian correlations resulting in non-significance in the frequentist analyses had Bayes Factors ranging from $BF_{10} = .382$ to $BF_{10} = 1.643$, which can be interpreted as 'anecdotical evidence', and are in agreement with the obtained results under the frequentist paradigm (Table 2).

When comparing scores by format and sample (see Table 3), results show an effect that can be attributable to the format on the MSES, evidencing statistically significant differences by format, with a large effect size (Cohen 1988). A post hoc analysis of the format effect found for the MSES, shows that this effect only appeared in the disengagement dimension ($F(1,30) = 13.339$; $p = .001$; $\eta_p^2 = .308$), with very strong evidence according to the Bayesian ANOVA ($BF_{10} = 31.099$), but not in the engagement dimension ($F(1,30) = .332$, $p = .569$, $\eta_p^2 = .011$; $BF_{10} = .227$). Thus, the scores reported in the disengagement dimension were significantly higher in the self-administered format for the total sample ($M = 44.99$; $SD = 14.71$) compared to those from the read-aloud format ($M = 37.32$; $SD = 10.67$) (Table 4).

In the remaining instruments, we did not observe interaction effects between the format and the group (i.e. LIF or TD), nor the main effects attributable to the format (Table 3). For example, as shown in Table 3,

Table 3. Comparison of participant scores by format and group.

Instrument[a]	Source	*F*	*dfn;dfd*	*p*	η_p^2	$BF_{(10)}$[b]	$BF_{(01)}$[c]
SEI	Format	1,934	*6;25*	*.114*	0.32	.720	1.390
	Group	1,113	*6;25*	*.383*	0.21	.896	1.116
	Group*Format	0.524	*6;25*	*.785*	0.11	.416	2.404
SPS	Format	.716	*4;27*	*.588*		.537	1.862
	Group	2.228	*4;27*	*.093*		.661	1.513
	Group*Format	.980	*4;27*	*.435*		.537	1.864
BMSLSS	Format	0.008	*1;30*	*.929*		.201	4.979
	Group	0.096	*1;30*	*.759*		.332	3.010
	Group* Format	0.022	*1;30*	*.884*		.090	11.145
MSES	Format	6.456	*2;29*	*.005*	.308	.227 EngS	4.406 EngS
						31.099 DisS	.032 DisS
	Group	.941	*2;29*	*.402*		.466 EngS	2.144 EngS
						.688 DisS	1.453 DisS
	Group* Format	0.773	*2;29*	*.471*		.120 EngS	8.323 EngS
						.975 DisS	1.026 DisS

[a]SEI: student engagement instrument; SPS: school participation scale; BMSLSS: brief multidimensional student life satisfaction scale; MSES: multidimensional school engagement scale (EngS: Engagement subscale; DisS: Disengagement subscale).
[b]Bayesian probability Factor of alternative over null hypothesis.
[c]Bayesian probability factor of null over alternative hypothesis.

Table 4. Estimation of the means by group and format application for MSES.

					95% CI	
Measurement	Group	Format	Mean	Dev. Error	Lower limit	Upper limit
Engagement	TD	Read-aloud	74.382	3.598	67.034	81.730
		Self-administered	74.800	2.361	69.979	79.621
	LIF	Read-aloud	69.334	3.380	62.432	76.236
		Self-administered	71.207	2.217	66.678	75.735
Disengagement	TD	Read-aloud	36.130	2.785	30.442	41.819
		Self-administered	41.056	3.735	33.427	48.684
	LIF	Read-aloud	38.367	2.616	33.024	43.710
		Self-administered	48.471	3.509	41.305	55.636

Note. LIF = low intellectual functioning group; TD = typical development group; CI = confidence interval.

the interaction effect between group and format for the Brief Multidimensional Student Life Satisfaction Scale (BMSLSS) shows strong evidence supporting the null model with a $BF_{01} = 11.145$. This result allows us to discard any interaction hypothesis between group and format for the BMSLSS, because the Bayes factor estimates that it is 11 times more likely that the data support the null model rather than the alternative one (i.e. interaction hypothesis). We can also discard with moderate evidence either a format ($BF_{01} = 4.979$) or a group effect ($BF_{01} = 3.010$) for this questionnaire.

Discussion

Students with LIF often experience barriers to participating in social research due to the literacy demands of the survey's typical self-administered format (Davies *et al.* 2017). These barriers limit their ability to share their opinions, attitudes, and perceptions related to any life situation (Davies *et al.* 2017), and therefore, exclude them from expressing their needs and experiences by themselves (Albuquerque 2021).

Although some scholars have stated the difficulties of the self-administered format for students with LIF and the suitability of the read-aloud format to measure math or reading, few studies have analyzed the impact of application formats on attitudes or opinion questionnaires for them. Moreover, according to the systematic review of Kooijmans *et al.* (2022) empirical evidence about the suitability of the self-administration format for people with LIF is scarce.

To the best of our knowledge, this is the first study that analyzes the effect of two application formats (i.e. read-aloud vs. self-administered) on psychological measures for LIF students compared to their TD peers using four instruments for psychoeducational research.

Regarding data quality, we found similar proportions of missing values in the self-administered condition compared to the read-aloud condition when assessing the LIF group. All these missing values were less than 2% in both formats, which is adequate (Tabachnick and Fidell 2013) and even better than the missing values found in the TD group.

We also obtained the same good quality of data for both self-assessment and read-aloud questionnaires, for LIF and TD groups, which is inconsistent with the results reported by Gresch *et al.* (2016) showing higher proportions of missing values in the self-administered condition compared to the read-aloud condition. In our view, three important aspects of our procedures may have contributed to obtain this result. First, we did not impose a time limit in any of the two conditions, and time limitation is an aspect that could raise missing

value rates, as reported by Gresch *et al.* (2016). Second, for both application formats, we asked the participants to answer all items, so this instruction could have led them to make an extra effort when answering the questionnaires, therefore reducing the omission rate. Third, the selected questionnaires did not have any of the seven aspects of question complexity considered by Redline *et al.* (2005). Therefore, they were rather simple to complete and did not contain any filter questions. These characteristics may have facilitated the questionnaires' comprehension for both LIF and TD students, and then made participants prone to omit very few items in both formats.

Regarding internal consistency, both formats showed adequate indices for the read-aloud and self-administered format for three out of four questionnaires (SPS, MSES, and BMSLSS), with no significant statistical differences between them (only the engagement subscale of MSES was significantly better in the read-aloud format, but it also had a good reliability index in the self-administered format). The obtained reliability indices for SPS in the read-aloud and self-administered formats, both in LIF and TD groups were similar to those reported by Pérez-Salas *et al.* (2019) in the validation sample. The same phenomenon was observed for the MSES, in which reliability indices for both formats and samples are similar to those reported in the validation sample (Wang *et al.* 2019). The BMSLSS also showed similar reliability indices in the formats and samples, as reported by Alfaro *et al.* (2015).

In the case of the SEI, most of the reliability indices obtained by the two samples for the two formats were lower than those reported in the validation sample in Chile (González *et al.* 2022). Although the read-aloud format condition showed slightly higher reliability indices than the self-administered one, these differences were not statistically significant (the only exception was the 'peer support at school' subscale for the TD group, where self-administration was statistically significantly better). Regarding this finding, we do not have a sound hypothesis that may explain why the SEI showed lower reliability coefficients than those observed for the validation sample. A plausible explanation is related to some specific psychometric properties of this instrument that may not work well with the particular characteristics of the present sample. Thus, it is necessary to further investigate and analyze the psychometric properties of this instrument, particularly its metric invariance in different populations.

Regarding convergent validity, the engagement dimension of MSES and almost all SPS subscales had appropriate intra-subject correlations between formats (self-administered and read-aloud) in the LIF group (as well as the TD group), suggesting that students had similar scores in both formats. This finding is important because the self-administered format is suitable for collective applications and is cost-effective. Hence, researchers could consider the inclusion of students with LIF in large research and compare results with TD groups in important psychoeducational measures like school engagement and school participation. Nevertheless, convergent validity for SEI, BMSLSS, and the disengagement dimensions of MSES was not good in the LIF sample, indicating that both formats could be evaluating different things in this population.

We detected a format effect for one instrument in one dimension: the disengagement dimension of the MSES scale. This finding demonstrates that in general, this dimension tends to show greater scores (i.e. more disengagement) in the self-administered format in both samples. These results could be influenced by the presence of an interviewer during the assessment because social desirability could affect students' responses when consulted about their behaviors, cognitions, and maladjusted emotions in school. Previous studies reported that participants with LIF have higher levels of socially desirable responses when answering questions with sensitive content (Langdon *et al.* 2010, Nelson and Liebel 2018). This has also been observed on students with TD, especially those with lower cognitive skills (Chang and Krosnick 2010). Regarding the items on the disengagement subscale, which ask about students' dysfunctional behaviors, social desirability can be expressed through the denial of socially undesirable characteristics or behaviors.

Remarkably, the absence of an interaction effect (i.e. format $\times$ group) or a format effect (i.e. self-administered vs. read-aloud) in all other questionnaires used in this study contradicts studies found in the literature (Lovett and Nelson 2021, Wood *et al.* 2018), which have stated that the read-aloud format could reduce the irrelevant variance of the measurement, and thus, improve test scores in students with disabilities, the so-called 'interaction hypothesis' (Sireci *et al.* 2005, Zuriff 2000). It also contradicts the 'differential boost hypothesis' (Fuchs and Fuchs 2001) that states that accommodation may be appropriate even if the TD students would benefit from it, as long as students with disabilities who need accommodation receive a significantly greater score increase.

These results are not surprising, given that the 'interaction hypothesis' (Sireci *et al.* 2005, Zuriff 2000) and 'the differential boost hypothesis' (Fuchs and Fuchs 2001) were originally designed for achievement questionnaires, not for attitude or opinion assessments like those used in our research. We believe it is crucial to emphasize this point and highlight the need for different criteria to evaluate the appropriateness of format changes for students with LIF when assessing psychological constructs. Therefore, we propose a new criterion to assess the validity of accommodations for instruments measuring attitudes or opinions, which can

serve as a guide for selecting the suitable format, either read-aloud or self-administered, for individuals with LIF. This criterion involves three aspects: (a) quality of data in both formats (i.e. a similar or better number of complete data in the chosen format); (b) equivalent reliability coefficients (i.e. no statistically significant differences between the reliability indexes of formats), and c) convergent validity (adequate correlation between formats). By adopting this criterion, we aim to establish a comprehensive framework for evaluating the suitability of format accommodations, thus contributing to more informed decisions when administering instruments to individuals with LIF.

We strongly consider that if both formats generate a good number of complete values, both have good reliability indices (without statistically significant differences between them); and both formats have a high correlation index, either application format could be used interchangeably depending on the needs of the researcher (i.e. convenience criterion based on equivalence).

On the other hand, if one of the two formats generates more complete data or shows higher reliability indices (with statistically significant differences between them), and there is a good correlation index between formats, that format should be the preferred 'format of choice' to assess people with LIF (i.e. quality criterion).

Of course, implementing the abovementioned suggestions requires that researchers conduct more studies of this kind to analyze the applicability of conventional instruments with both formats for people with LIF. However, it is important to mention that there are challenges related to the difficulty of accessing individuals with LIF to carry out studies of this type. We believe that the use of Bayesian statistics can help us to circumvent the difficulty of sample sizes and contribute to generating empirical evidence regarding the equivalence of traditional instruments to be applied to people with LIF and favor their inclusion in social research.

There are some limitations that arise when interpreting our results. First, we only included students with mild cognitive impairment in the LIF group; however, the present study does not offer evidence on the read-aloud accommodation's effect on severe intellectual disability. Second, there were no constraints on the time required by the participants to complete a questionnaire; therefore, self-administered questionnaires that impose time-limited responses could generate lower-quality data than those reported here, as time constraints could lead to more omitted or quicker responses.

For the specific case of students with LIF, our results support our belief that both questionnaire formats: read-aloud, and self-administered with simple questions, could be appropriate alternatives to assess students' attitudes or opinions, and they may increase access to evaluations in the field of social research. However, it is important to evaluate possible format effects for different questionnaires as differences could arise as we have put in evidence here.

Furthermore, it results necessary to conduct further studies on the effect of the read-aloud accommodation on the evaluation of attitudes of students with disabilities other than those included in the present study, such as students with severe ID, motor, hearing, or visual impairments.

Finally, it would be also important to explore the effect of the researcher's presence in the read-aloud format and the bias of social desirability, specifically for questions which imply sensitive content.

By developing a deeper understanding of the reliability and validity of psychological assessment for LIF students, the present study contributes to providing researchers, school systems, and policymakers with empirical evidence to facilitate access and support to the participation of all students in social research.

Note

1. Following to Lees-Warley and Rose (2015), we use the umbrella term 'low intellectual functioning' to refer to individuals that have an intelligence quotient below 85, that is those that have mild intellectual disability (IQ $<$ 70) or borderline intellectual functioning (IQ $<$ 85).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work received funding from the National Research and Development Agency of the Chilean Government [ANID/CONICYT, Proyecto FONDECYT Regular 1181265].

Data availability statement

The data that support the findings of this study are available from the corresponding author (CPS) upon reasonable request.

References

Albuquerque, C. 2021. Needs of older people with intellectual disabilities: Variables influencing inter-respondent (client vs staff) agreement. *International Journal of Developmental Disabilities*, 69, 256–264.

Alfaro, J., Guzmán, J., García, C., Sirlopú, D., Gaudlitz, L. and Oyanedel, J. 2015. Propiedades psicométricas de la Escala Breve Multidimensional de Satisfacción con la Vida para Estudiantes (BMSLSS) en población infantil chilena (10–12 años). *Universitas Psychologica*, 14, 29–42.

Andreou, G., Athanasiadou, P. and Tzivinikou, S. 2019. Accommodations on reading comprehension assessment for students with learning disabilities: A review study. *Psychology*, 10, 521–538.

Appleton, J. J., Christenson, S. L., Kim, D. and Reschly, A. L. 2006. Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of School Psychology*, 44, 427–445.

Bowling, A. 2005. The mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, 27, 281–291.

Buzick, H. and Stone, E. 2014. A meta-analysis of research on the read-aloud accommodation. *Educational Measurement*, 33, 17–30.

Buzick, H. M. 2019. Testing accommodation and the measurement of student academic growth. *Educational Assessment*, 24, 57–72.

Casas, F., Sarriera, J. C., Abs, D., Coenders, G., Alfaro, J., Saforcada, E. and Tonon, G. 2012. Subjective indicators of personal well-being among adolescents. performance and results for different scales in Latin-language speaking countries: A contribution to the international debate. *Child Indicators Research*, 5, 1–28.

Chang, L. and Krosnick, J. A. 2010. Comparing oral interviewing with self-administered computerized questionnaires: An experiment. *Public Opinion Quarterly*, 74, 154–167.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale: Routledge.

Davies, D. K., Stock, S. E., King, L., Wehmeyer, M. L. and Shogren, K. A. 2017. An accessible testing, learning and assessment system for people with intellectual disability. *International Journal of Developmental Disabilities*, 63, 204–210.

De Leeuw, E. 1992. *Data Quality in Mail, Telephone and Face to Face Surveys*. Amsterdam: T. T. Publikaties.

Diedenhofen, B. and Musch, J. 2016. Cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science*, 11, 51–60.

Dienes, Z. 2011. Bayesian versus Orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274–290.

Emerson, E., Felce, D. and Stancliffe, R. J. 2013. Issues concerning self-report data and population-based data sets involving people with intellectual disabilities. *Intellectual and Developmental Disabilities*, 51, 333–348.

Finlay, W. M. and Lyons, E. 2001. Methodological issues in interviewing and using a self-report questionnaire with people with mental retardation. *Psychological Assessment*, 13, 319–335.

Fuchs, L. S. and Fuchs, D. 2001. Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research & Practice*, 16, 174–181.

Giusto, M. and Ehri, L. C. 2019. Effectiveness of a partial read-aloud test accommodation to assess reading comprehension in students with a reading disability. *Journal of Learning Disabilities*, 52, 259–270.

Gresch, C., Strietholt, R., Kanders, M. and Solga, H. 2016. Reading-aloud versus self-administered student questionnaire: An experiment on data quality. In: H. P. Blossfeld, J. Von Maurice, M. Bayer and J. Skopek, eds. *Methodological issues of longitudinal surveys: The example of the national education panel survey*. Wiesbaden: Springer, pp.561–578.

Goegan, L. D., Radil, A. I. and Daniels, L. M. 2018. Accessibility in questionnaire research: Integrating universal design to increase the participation of individuals with learning disabilities. *Learning Disabilities*, 16, 177–190. https://files.eric.ed.gov/fulltext/EJ1194555.pdf

González, L., Oñate, V., Longos, M., Espinoza, L., Lema, C., Pérez-Salas, C. P. and Sáez-Delgado, F. 2022. Análisis psicométrico del instrumento de compromiso escolar (SEI) en estudiantes secundarios en Chile. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica*, 65, 47–60.

Jaynes, E. T. 2003. *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Jeffreys, H. 1984. *Theory of probability*, 3rd ed. New York: Oxford University Press.

John-Akinola, Y. O. and Nic-Gabhainn, S. 2014. Children's participation in school: A cross-sectional study of the relationship between school environments, participation and health and well-being outcomes. *BMC Public Health*, 14, 964–974.

Kass, R. E. and Raftery, A. E. 1995. Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Kooijmans, R., Mercera, G., Langdon, P. E. and Moonen, X. 2022. The adaptation of self-report measures to the needs of people with intellectual disabilities: A systematic review. *Clinical Psychology*, 29, 250–271.

Langdon, P. E., Clare, I. C. and Murphy, G. H. 2010. Measuring social desirability amongst men with intellectual disabilities: The psychometric properties of the self-and other-deception Questionnaire—Intellectual disabilities. *Research in Developmental Disabilities*, 31, 1601–1608.

Lee, M. D. and Wagenmakers, E. J. 2014. *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lees-Warley, G. and Rose, J. 2015. What does the evidence tell us about adults with low intellectual functioning who deliberately set fires? A systematic review. *International Journal of Developmental Disabilities*, 61, 242–256.

Li, H. 2014. The effects of read-aloud accommodation for students with and without disabilities: A meta-analysis. *Educational Measurement*, 33, 3–16.

Lovett, B. J. and Nelson, J. M. 2021. Systematic review: Educational accommodations for children and adolescents with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 60, 448–457.

Lozano, F., Lobos, J. M., March, J. R., Carrasco, E., Barros, M. B. and González-Porras, J. R. 2016. Self-administered versus interview-based questionnaires among patients with intermittent claudication: Do they give different results? A cross-sectional study. *Sao Paulo Medical Journal = Revista Paulista de Medicina*, 134, 63–69.

Mineduc. 2009. *Orientaciones técnicas para la evaluación diagnóstica de estudiantes que presentan necesidades educativas especiales asociadas a discapacidad intelectual*. Santiago: Ministerio de Educación de Chile.

Nelson, J. M. and Liebel, S. W. 2018. Socially desirable responding and college students with dyslexia: Implications for the assessment of anxiety and depression. *Dyslexia*, 24, 44–58.

Ortega, A. and Navarrete, G. 2017. Bayesian hypothesis testing: An alternative to null hypothesis significance testing (NHST) in psychology and social sciences. In: J. Prieto, ed. *Bayesian inference*. Rijeka: InTech, pp.235–254.

Ortega, A., Wagenmakers, E.-J., Lee, M. D., Markowitsch, H. J. and Piefke, M. 2012. A Bayesian latent group analysis for detecting poor effort in the assessment of malingering. *Archives of Clinical Neuropsychology*, 27, 453–465.

Pérez-Salas, C. P., Sirlopú, D. and Cobo, R., and A. A. 2019. Análisis bifactorial de la escala de participación escolar en una muestra de estudiantes chilenos. *Revista Iberoamericana de Diagnóstico y Evaluación Psicológica*, 52, 27–39.

Pérez-Salas, C. P. 2021. *Psychometric properties of multidimensional scale of school engagement*. Unpublish manuscript. Departamento de Psicología, Universidad de Concepción: Concepción.

Redline, C. D., Dillman, D. A., Carley–Baxter, L. and Creecy, R. H. 2005. Factors that influence reading and comprehension of branching instructions in self–administered questionnaires. *Allgemeines Statistisches Archiv*, 89, 21–38.

Rogers, C. M., Thurlow, M. L., Lazarus, S. S. and Liu, K. K. 2019. *A summary of the research on effects of test accommodations: 2015–2016 (NCEO Report 412)*. University of Minnesota, National Center on Educational Outcomes.

Seligson, J. L., Huebner, E. S. and Valois, R. F. 2003. Preliminary validation of the multidimensional brief study' life satisfaction scale (BMSLSS). *Social Indicators Research*, 61, 121–145.

Sireci, S. G., Scarpati, S. and Li, S. 2005. Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75, 457–490.

Sireci, S. G., Banda, E. and Wells, C. S. 2018. Promoting valid assessment of students with disabilities and english learners. In: S. Elliott, R. Kettler, P. Beddow, A. and Kurz, eds. *Handbook of accessible instruction and testing practices*. New York: Springer. pp.231–246.

Shogren, K. A., Bonardi, A., Cobranchi, C., Krahn, G., Murray, A., Robinson, A. and Havercamp, S, 2021. State of the field: The need for self-report measures of health and quality of life for people with intellectual and developmental disabilities. *Journal of Policy and Practice in Intellectual Disabilities*, 18, 286–295.

Spiel, C. F., Mixon, C. S., Holdaway, A. S., Evans, S. W., Harrison, J. R., Zoromski, A. K. and Yost, J. S. 2016. Is reading tests aloud an accommodation for youth with or at risk for ADHD? *Remedial and Special Education*, 37, 101–112.

Tabachnick, B. G. and Fidell, L. S. 2013. *Using multivariate statistics*. Boston: Pearson.

Team JASP. 2018. JASP (Version 0.8.4), Computer Software.)

Thurlow, M. L., Lazarus, S. S. and Hodgson, J. R. 2012. Leading the way to appropriate selection, implementation, and evaluation of the read-aloud accommodation. *Journal of Special Education Leadership*, 25, 72–80.

Wang, M.-T., Fredricks, J., Ye, F., Hofkens, T. and Linn, J. S. 2019. Conceptualization and assessment of adolescents' engagement and disengagement in School: A multidimensional school engagement scale. *European Journal of Psychological Assessment*, 35, 592–606.

Wetzels, R., Raaijmakers, J. G., Jakab, E. and Wagenmakers, E. J. 2009. How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, 16, 752–760.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J. and Wagenmakers, E. J. 2011. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6, 291–298.

Wood, S. G., Moxley, J. H., Tighe, E. L. and Wagner, R. K. 2018. Does the use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis. *Journal of Learning Disabilities*, 51, 73–84.

Zuriff, G. E. 2000. Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education*, 13, 99–117.

Appendix. Evaluation of counterbalance effect

Instrument[a]	Analysis	Source	F	$df_n;df_d$	*p*
SEI	Mixed MANOVA of repeated measurements	Position*Format	.649	6;25	.690
		Position	2.217	6;25	.075
		Format	2.2	6;25	.077
SPS	Mixed MANOVA of repeated measurements	Position*Format	1.920	4;27	.136
		Position	1.107	4;27	.374
		Format	.833	4;27	.516
BMSLSS	Mixed ANOVA of repeated measurements	Position*Format	2.562	1;30	.120
		Position	.699	1;30	.410
		Format	.047	1;30	.829
MSES	Mixed MANOVA of repeated measurements	Position*Format	2.394	2;29	.109
		Position	.826	2;29	.448
		Format	8.724	2;29	.001

[a]SEI: student engagement instrument; SPS: school participation scale; BMSLSS: brief multidimensional student life satisfaction scale; MSES: multidimensional school engagement scale.