European Journal of Psychological Assessment

To Guess or Not to Guess? – A Study of the Impact of Guessed Answers of the Adaptive Behavior Assessment System, Third Edition (ABAS-3)

Sofia Santisi, Matteo Ciancaleoni, Michele Borghetto, Rosa Ferri, Maria Landi, and Margherita Lang

Online First Publication, June 12, 2025. https://dx.doi.org/10.1027/1015-5759/a000905

CITATION

Santisi, S., Ciancaleoni, M., Borghetto, M., Ferri, R., Landi, M., & Lang, M. (2025). To guess or not to guess? —A study of the impact of guessed answers of the Adaptive Behavior Assessment System, Third Edition (ABAS-3). *European Journal of Psychological Assessment*. Advance online publication. https://dx.doi.org/10.1027/1015-5759/a000905

Multistudy Report

To Guess or Not to Guess? – A Study of the Impact of Guessed Answers of the *Adaptive Behavior Assessment System*, Third Edition (ABAS-3)

Sofia Santisi[1], Matteo Ciancaleoni[2], Michele Borghetto[3], Rosa Ferri[4], Maria Landi[4], and Margherita Lang[5]

[1]Department NEUROFARBA, University of Florence, Florence, Italy
[2]Hogrefe Editore, Florence, Italy
[3]Azienda Unità Locale Socio Sanitaria n. 2 Marca trevigiana, Treviso, Italy
[4]Department of Dynamic, Clinical and Health Psychology, University of Rome La Sapienza, Rome, Italy
[5]Association for Research in Clinical Psychology, Milan, Italy

Abstract: The Adaptive Behavior Assessment System, Third Edition (ABAS-3) is widely recognized as a valuable tool for assessing adaptive behavior. In addition to indicating how frequently a certain behavior is displayed, respondents may also specify if they guessed their answer. Despite the author's suggestions, it has not yet been empirically established if the presence of guessed answers represents a discriminant variable to continue scoring in the adaptive skill area that presents four or more guessed answers. Therefore, the present study aimed to establish if there are differences between protocols with guessed and no guessed answers in ABAS-3. To achieve this goal three different studies were conducted for Parent, Teacher, and Adult Forms. The Item Response Theory approach was adopted to assess the Differential Item Functioning (DIF) between participants who had given no guessed answer (NGG) and those who had given at least one guessed answer (GG). In the subscales with DIF items, these constituted less than 10% of the total items, or DIF's effect size was negligible, therefore, all the subscales of ABAS-3 were invariant with respect to membership in NGG and GG. The present study makes an important contribution in interpreting ABAS-3 profiles with guessing answers.

Keywords: adaptive behavior assessment system, ABAS-3, invariance, differential item functioning, item response theory

The Adaptive Behavior Assessment System, Third Edition (ABAS-3; Harrison & Oakland, 2015) is widely recognized as a valuable tool for assessing adaptive behavior across different domains and age groups available to health and education professionals (Mecca et al., 2022). ABAS-3 reflects current standards for describing adaptive behavior and diagnosing conditions in which it may be impaired (e.g., American Psychiatric Association, 2022; Schalock et al., 2012; World Health Organization, 2007), promoting assessment of adaptive behavior to target daily living skills that may be improved through intervention, thus enhancing an individual's quality of life. ABAS-3 can be used with individuals from birth to 89 years in many settings (e.g., schools, clinics, hospitals, residential facilities, and community agencies).

ABAS-3 has been adapted into Dutch (Kreemers et al., 2020), Arabic from Oman (Emam et al., 2020), Danish (Harrison & Oakland, 2020), Swedish (Zander & Bölte, 2020), Portuguese from Brazil (Mecca et al., 2022), and Italian (Ferri et al., 2025).

ABAS-3 measures adaptive behavior at three different levels. The highest level is the General Adaptive Composite (GAC), which is composed of all measured ability areas and thus provides an overall estimate of adaptive behavior. At the next level there are three adaptive domains, each of which includes multiple individual skill areas: *Conceptual* (behaviors necessary to communicate with others, apply academic skills, and manage and complete tasks), *Social* (behaviors necessary to engage in interpersonal interactions, act with social responsibility, and use leisure time), and *Practical* (behaviors necessary to meet personal and health needs, to care for the home, classroom or work environment, and to function in a community). Finally, at the lowest level, there are specific adaptive skills, divided into 11 scales that vary according to age. Specifically, the *Conceptual* domain is assessed through the subscales (1) *Communication*,

(2) *Functional Academics/Pre-Academics,* and (3) *Self-Direction*; the *Social* domain is evaluated using (4) *Leisure* and (5) *Social* subscales; finally, the *Practical* domain is examined through the subscales (6) *Community Use*, (7) *Home/School Living*, (8) *Health and Safety*, (9) *Self-Care*, (10) *Work* (for young adults), and (11) *Motor* (for young children).

ABAS-3 consists of five assessment forms to collect information about an individual's adaptive behaviors in relevant contexts. These age-specific forms are differentiated for parents (Parent/Primary Caregiver Form from 0 to 5 years and Parents Form from 5 to 21 years), teachers (Teacher/Daycare Provider Form from 2 to 5 years and Teacher Form from 5 to 21 years), and adults (Adult Form from 16 to 89 years; this form may also be used by adults to perform a self-rating).

Respondents rate ABAS-3 items by indicating whether the individual can independently display behavior and, if so, how frequently the behavior is displayed when necessary, ranging from 0 to 3 points (0 = *Is not able*; 1 = *Never*; 2 = *Sometimes;* 3 = *Almost always*).

Respondents also may specify a rating based on more limited information, indicating that they guessed or estimated it by checking the box in the "Check only if you guessed" column next to each item (Harrison & Oakland, 2015).

A potential limitation of the ABAS-3 concerns the number of guessed answers. According to the authors' suggestions, occasional guessing is acceptable, but if there are four or more guessed answers in an adaptive skill area, it must be discussed with the respondent whether some guessed answers can be legitimately eliminated. If four or more guesses remain, the professional decides whether to continue assessing the adaptive skill area (Harrison & Oakland, 2015). This indication could be applied when the ABAS-3 is administered in person, allowing the administrator to ask questions to reduce the number of guessed answers. However, especially with online and remote administration, this is more rarely done today. Although online administration should follow the same methodological principles as in-person interviews, ensuring consistent guidelines and quality control, one possible consequence of this administration method could be an increase in the number of guessing answers.

Despite that, there are no studies that investigated the impact of guessing answers to the ABAS-3 items. Shedding light on this problem is particularly important to clarify the role of guessing answers as a discriminant variable in determining whether to continue scoring the adaptive skill area or not in the presence of four or more guessed answers in an adaptive skill area. This would generally improve the quality of adaptive behavior assessment, with important consequences for people's lives.

The present study aimed to evaluate the test invariance across protocols with guessed answers and protocols with no guessed answers in the Italian version of ABAS-3 to reduce the potential limitation of the ABAS-3 instrument due to the large number of guessed answers.

To achieve this goal, the Item Response Theory (IRT) approach was adopted to assess the Differential Item Functioning (DIF; Nering & Ostini, 2011), an efficient method for analyzing the measurement invariance of a test (Greiff & Scherer, 2018) that makes possible to determine whether trait scores across groups are comparable and have the same meaning across groups (Reise et al., 1993).

According to the authors' criteria (Harrison & Oakland, 2015), in the presence of four or more guessed answers, the professional decides to maintain the profile as valid or invalid. Despite that, no studies have been carried out to empirically determine the number of guessing answers that makes the profile invalid; moreover, it is reasonable to think that this cutoff may be different among different versions and scales of ABAS-3. For this reason, the amount of biased items has been evaluated to empirically determine whether profiles with a large amount of guessed answers could be considered valid. Since the main goal was to establish the amount of guessing answers that makes the profile invalid for each version and each scale, to empirically evaluate the presence and the impact of biased items, in this study two groups were created: a no-guessing group (NGG), consisting of participants who had given no guessed the answer, and a guessing group (GG), composed by participants who had given at least one guessed answer. In this way, it was possible to compare the IRT parameters of the items and to assess the presence of DIF between the "perfect" group without guessed answers and the other participants. This procedure made it possible to investigate the number of biased items for each scale in all versions which make the profile invalid due to the presence of large-scale-level bias. The distribution of guessed answers for each subscale of each form of ABAS-3 is shown in Table 1.

To evaluate the presence of DIF in the Italian ABAS-3 forms, three studies were conducted: in Study 1, DIF between NGG and GG in the two parent forms was assessed, in Study 2, in the two teacher forms, and Study 3 in the adult form.

Study 1

Parent/Primary Caregiver Form (ages 0–5 years, 11 months) measures the adaptive functioning of infants, toddlers, and preschoolers at home and in other settings. Parent Form (ages 5–21 years, 11 months), on the other hand, measures the adaptive functioning of older children and

Table 1. Mean (*M*) and standard deviation (*SD*) of the guessed answers for each subscale of each form of ABAS-3

	Parent/Primary caregiver form		Teacher/Daycare provider form		Parent form		Teacher form		Adult form	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Communication	0.53	2.44	0.42	1.19	1.40	4.48	1.78	2.96	1.13	3.44
Community Use	0.71	2.29	–	–	1.69	4.53	4.79	3.96	1.64	4.10
Functional Academics/Pre-Academics	0.46	1.97	0.95	1.75	1.37	4.01	3.83	3.93	1.84	3.96
Home Living	0.37	1.59	–	–	0.95	3.94	–	–	1.72	4.47
School Living	–	–	0.23	0.70	–	–	2.33	2.97	–	–
Health and Safety	0.60	1.82	1.27	1.86	1.20	3.28	3.06	3.04	1.28	3.18
Leisure	0.49	1.88	0.86	1.70	0.92	3.24	3.22	4.17	1.54	3.90
Self-Care	0.31	1.66	0.18	0.65	0.71	3.75	2.11	3.12	0.77	3.35
Self-Direction	0.57	1.95	0.45	0.99	1.35	4.01	1.39	2.21	1.79	4.27
Social	0.50	1.70	0.51	1.14	1.39	4.20	3.64	4.73	1.23	3.66
Motor	0.46	1.99	0.97	1.80	–	–	–	–	–	–
Work	–	–	–	–	1.24	4.50	0.55	2.69	2.80	6.30

young adults in the home and community. Both forms can be completed by parents or other primary care providers (Harrison & Oakland, 2015). The aim of Study 1 was to evaluate the invariance of the two Parents' forms across NNG and GG groups; for that reason, the DIF analyses using IRT in ABAS-3 items of each subscale were performed.

Materials and Methods

Participants

A total of 1,469 parents took part in Study 1: in detail, parents of 394 children (M_{age} = 2.48 years; *SD* = 1.39; male = 50.51%) filled out the Parent/Primary Caregiver Form and parents' of 1,075 children and young adults (M_{age} = 12.63 years; *SD* = 4.24; male = 51.44%) filled out the Parent Form. Based on the presence of guessed answers given, participants were divided into NGG and GG: for each subscale, the NGG group was composed of people with no guessed answer in all subscale items, while the GG group was formed by parents who guessed at least one item in the target subscale. The sample size was determined based on the sample used in the original version of the ABAS-3 and demographic data from the Italian National Institute of Statistics, aiming to ensure representativeness. No inclusion or exclusion criteria were applied to maximize the representativeness of the sample.

Procedure

The first step was to evaluate the possibility of using IRT models, so IRT assumptions were investigated. To test the presence of local dependence, the LD statistic (Chen & Thissen, 1997) was used; this assumption requires that for each latent trait level, the items must be independent of each other: this assumption is met when the LD values are less than 10.

Confirmatory factor analysis (CFA) was performed for each subscale of the ABAS-3 to evaluate the unidimensionality assumption: Comparative Fit Index (CFI; Bentler, 1990), Tucker–Lewis Index (TLI; Tucker & Lewis, 1973), and the Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980) were used to test the model fit. For the TLI and CFI indices, values above .90 are indicative of acceptable fit, while values above .95 of excellent fit (Hu & Bentler, 1999). The RMSEA value is acceptable when it is below .08 and good when it is below .05 (Kline, 2010). Moreover, the strength of the relationship between the factor and the measured variable was analyzed through the factor-loadings (λ) of the items, and the minimum value required was .30 (Kline, 1998).

Once the IRT assumptions were met, IRT model was applied. In detail, since the ABAS-3 response scale is polytomous, the graded response model (GRM; Samejima, 1969) was used.

To test the adequacy of the model, each item's fit under the GRM was tested by computing the S-χ^2 statistics. Given that using larger samples results in a greater likelihood of significant chi-square differences, item fit was investigated by taking false positives into account, and the critical *p*-value of .001 was employed. Because of GRM estimates for each item, the discrimination parameter and a number of location parameters equal to the number of response alternatives minus one, analyses of both uniform (location parameters) and nonuniform (discrimination parameter) DIF across NGG and GG were performed. DIF investigation is an iterative procedure and in the first step (purification phase), the aim is to identify the DIF candidate's items, that is, items that show a potential bias during the purification phase: in this phase, the anchor items are selected

through a process of *log-likelihood* comparisons; in the second step, all items in the subscale that showed no DIF in the first phase are used as anchor items to repeat the DIF procedure to the candidate items to finally evaluate the presence of DIF in these candidate items. During this iterative process, the DIF status of the items may change. To evaluate the presence of DIF, in both phases, the chi-square statistic was used. Because large samples lead to a greater likelihood of significant chi-square differences, the critical value at $p = .001$, rather than the one at $p = .05$, was used (Stone & Zhang, 2003).

To determine if the detected DIF was meaningful (Kim et al., 2007), the DIF effect size was analyzed using the Expected Score Standardized Difference (ESSD; Meade, 2010): this index can be interpreted following Cohen's *d* criteria, whereby: $d = .20$ corresponds to a small effect size; $d = .50$ to a medium effect size; $d = .80$ to a large effect size (Cohen, 1988).

To evaluate the impact of DIF in the whole subscale, two different approaches were applied. The first approach is based on the number of DIF items in a subscale: subscales whose items with DIF accounted for less than 10% of the total items were considered invariant concerning NGG or GG membership (Budgell et al., 1995). Moreover, the effect size of the Differential Test Functioning (DTF) using the Expected Test Score Standardized Difference (ETSSD; Meade, 2010) was evaluated: this index can be interpreted as a Cohen's *d*.

In addition, a graphical comparison between the Test Information Functions (TIFs) of the NGGs and GGs of subscales with items with DIF was made to assess the measurement precision of the test at different levels of the measured trait (Embretson & Reise, 2013; Nering & Ostini, 2011) in both groups.

All analyses were performed using RStudio (Mirt package; Chalmers, 2012) software.

Results

Results attested that the IRT assumptions were met. In detail, local independence in each subscale of the two parent forms was attested because no LD value was found to be greater than 10.

CFA analyses confirmed the unidimensionality of each subscale (Table 2); all models showed a good fit except for the Leisure and Work subscales of the Parent Form: unidimensionality of the Leisure scale was approximately confirmed due to the low values of CFI and TLI indices and the high value of the RMSEA; minor issues were found in the Work subscale that presents only a small elevation in the RMSEA index. In addition, factor loadings were all adequate (>.30) with minor exceptions for Communication (items 1, 2, and 3), Health and Safety (items 1 and 2), and Social (item 1) subscales of the Primary Parent/Caregiver Form. The unidimensionality assumption was met, with only minor doubts concerning the Leisure subscale of the Parent Form: these doubts were not so consistent as to prevent the continuation of the analyses; for that reason, the next step was the investigation of item fit under GRM.

Regarding items' fit, all items of both forms showed a non-significant $S\text{-}\chi^2$ value, indicating a good fit.

Results of the purification phase showed that almost all the items of each subscale of the two forms had no DIF. As shown in Table 3, the only exceptions were found in the Communication (items 26), Home Living (item 15), and Social subscales (item 11) of the Primary Parent/Caregiver Form and in the Community Use subscale (item 23) of the Parent Form, where uniform DIF was found. Moreover, item 11 of the Social subscale of the Primary Parent/Caregiver Form presented non-uniform DIF. In the second step, all other items were used as anchor items, and the DIF analyses were carried out. After the iterative process of DIF analysis, only item 15 in the Home Living subscale and item 26 in the Communication subscale of the Parent/Primary Caregiver Form continued to show uniform DIF (Table 3).

The DIF effect size of item 15 of the Home Living was small to medium (ESSD = .365), whereas the DIF effect size of item 26 of the Communication subscale was negligible (ESSD = .062). Therefore, since each subscale of both parent forms of ABAS-3 had less than 10% of items with DIF or no items with DIF, both can be considered invariant concerning membership in NGG and GG.

Further confirming the invariance of the instrument, the DTF effect size was negligible in both Home Living (ETSSD = .020) and Communication (ETSSD = .002) subscales.

According to these results, the TIFs of NGG and GG, computed in the two subscales in which there was one DIF item, were almost perfectly overlapping (Figure 1), demonstrating how the functioning of the two subscales was essentially the same in the two groups.

Discussion

In the first study, the invariance of the two parent forms of the Italian version of the ABAS-3 across parents who guessed and no-guessed answers was examined.

After verifying the IRT assumptions, such as item local independence and the unidimensionality of each subscale of the two versions, DIF analyses applying GRM were performed.

Only two items with DIF were found: in detail, no DIF items were identified in the Parent form, while in the Home Living and Communication subscales of the Primary Parent/Caregiver Form there was one item with a differential functioning across the two subgroups that differed for the

Table 2. CFA fit indices of the two parent forms of ABAS-3

	Parent/Primary caregiver form			Parent form		
Subscale	RMSEA	CFI	TLI	RMSEA	CFI	TLI
Communication	.035, CI 90% [0.027, 0.042]	.997	.996	.053, CI 90% [0.049, 0.057]	.961	.957
Community Use	.049, CI 90% [0.040, 0.057]	.985	.983	.054, CI 90% [0.050, 0.058]	.984	.982
Functional Academics/Pre-Academics	.072, CI 90% [0.065, 0.079]	.975	.972	.046, CI 90% [0.042, 0.050]	.990	.991
Home Living	.067, CI 90% [0.060, 0.074]	.989	.965	.068, CI 90% [0.065, 0.072]	.956	.951
Health and Safety	.067, CI 90% [0.060, 0.075]	.974	.971	.042, CI 90% [0.037, 0.047]	.989	.964
Leisure	.031, CI 90% [0.021, 0.040]	.995	.994	.090, CI 90% [0.085, 0.094]	.866	.845
Self-Care	.032, CI 90% [0.023, 0.041]	.994	.994	.044, CI 90% [0.041, 0.048]	.977	.974
Self-Direction	.063, CI 90% [0.056, 0.070]	.979	.976	.060, CI 90% [0.056, 0.064]	.956	.951
Social	.036, CI 90% [0.027, 0.045]	.995	.995	.068, CI 90% [0.064, 0.071]	.945	.940
Motor/Work	.022, CI 90% [0.007, 0.032]	.998	.998	.082, CI 90% [0.064, 0.098]	.946	.938

Table 3. Items with uniform DIF (b) and non-uniform DIF (a) according to NGG or GG group membership in the parent forms of ABAS-3

			Purification phase						DIF analysis							
			*a*DIF			*b*DIF			*a*DIF			*b*DIF			DIF effect size	DTF effect size
Form	Subscale	Item	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*	ESSD	ETSSD
Parent/Primary	Home Living	15	–	–	–	18.88	3	< .001	–	–	–	19.24	3	< .001	.365	.020
Caregiver Form	Communication	26	–	–	–	23.43	3	< .001	–	–	–	20.37	3	< .001	.062	.002
	Social	11	12.72	1	< .001	18.81	3	< .001	2.06	1	.151	8.15	3	.043	–	–
Parent Form	Community Use	23	–	–	–	19.03	3	< .001	–	–	–	14.71	3	.002	–	–

presence of guessed answers. In both cases, items with DIF were less than 10% of the total number of subscale items. In the former case, a small-to-medium effect size of DIF was found, and in the latter case, the effect size was negligible. The invariance of the whole subtests was confirmed by effect size indices and by the inspection of the Test Information Functions.

Therefore, it is possible to consider the Parent Forms as invariant measures of adaptive behavior concerning the presence of guessing answers.

Study 2

Teacher/Daycare Provider Form (ages 2–5 years, 11 months) measures the adaptive functioning of toddlers and preschoolers in a daycare center, home daycare, preschool, or school setting. Teacher Form (ages 5–21 years, 11 months), on the other hand, measures the adaptive functioning of older children and young adults in a school setting. Both forms can be completed by teachers, teacher assistants, or other daycare, childcare providers, or school personnel (Harrison & Oakland, 2015).

The aim of Study 2 was to evaluate the invariance of the two Teachers' forms across NNG and GG groups; for that reason, DIF analyses using IRT in ABAS-3 items of each subscale were performed.

Materials and Methods

Participants

A total of 1,085 teachers took part in Study 2: in detail, teachers of 344 children (M_{age} = 44.83 months; *SD* = 12.95; male = 47.09%) completed the Teacher/Daycare Provider Form and teachers of 683 children and young adults (M_{age} = 12 years; *SD* = 3.66; male = 52.542%) completed the Teacher Form. Based on the number of guessed answers given, participants were divided into NGG and GG using the same criteria as in Study 1.

Procedure

The procedure followed is the same as in Study 1.

Results

Results attested that the LD assumption was met: no LD value was found to be greater than 10 in each subscale of the two teacher forms. Differently, the unidimensionality assumption was not met for all scales (Table 4): in particular, the Health and Safety and Self-Direction scales of the Teacher/Daycare Provider Form showed poor fit indices; moreover, for School Living, Leisure, Self-Care, and Social, RMSEA values were higher than the cutoff. Minor issues were found in the Teacher Form: Leisure and Social subscales showed slightly lower TLI values and not good values of the RMSEA, while Self-Care and Self-Direction subscales

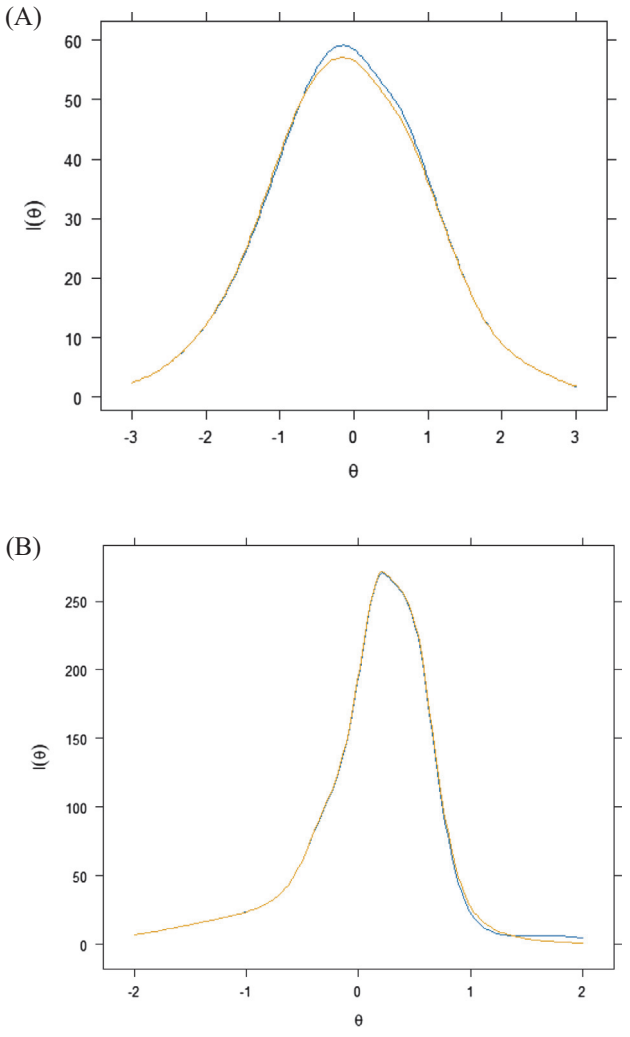

Figure 1. TIFs of NGG (blue line) and GG (orange line) of the two subscales of the primary parent/caregiver form composed by items with DIF. (A) Home Living; (B) Communication.

showed unacceptable RMSEA values. Despite that, the factor loadings were all adequate (>.30), with minor exceptions for Communication (item 26), Health and Safety (item 4), and Self-Care (item 2) subscales of the Teacher/Daycare Provider Form. Therefore, despite knowing that the unidimensionality assumption was not fully met for all subscales, especially for the Health and Safety and Self-Direction subscales of the Teacher/Daycare Provider Form, the next step was the investigation of item fit under GRM in order to test the presence of biased item due to the relevance of this purpose. All items of both forms showed a non-significant $S\text{-}\chi^2$ value, indicating a good fit, with a minor exceptions in the Motor subscale of the Teacher/Daycare Provider Form (items 1, 3, and 11).

Results of the purification phase showed that only a few items of the two teacher forms had potential DIF (Table 5). In the second step, all non-biased items were used as anchor items and DIF analysis was carried out. After the iterative process of DIF analysis, only item 22 of the Functional Academics subscale of the Teacher Form continued to show non-uniform DIF, only item 11 of the School Living subscale of the Teacher Form no longer showed uniform DIF, whereas all the candidate items of the Teacher/Daycare Provider Form, that belong to the no-unidimensional scales, continued to show uniform DIF (Table 5).

In the Teacher Form, DIF effect sizes of item 12 of the Health and Safety subscale (ESSD = .141), item 20 of School Living subscale (ESSD = .275) and item 6 of Leisure subscale (ESSD = .129) were small, whereas DIF's effect sizes of item 22 of Functional Academics subscale (ESSD = .496) and item 18 of School Living subscale (ESSD = .625) were medium. In the Teacher/Daycare Provider Form, DIF effect size of item 10 of the Health and Safety subscale was medium (ESSD = .492), whereas DIF effect size of item 15 of the Health and Safety subscale (ESSD = .964) was large, as well as DIF effect sizes of items 8 (ESSD = .839) and 23 (ESSD = 1.019) of the Self-Direction subscale. Finally, the DIF effect size of item 17 of the Self-Direction subscale (ESSD = .703) was medium to large. Therefore, since each subscale of both teacher forms of ABAS-3 had less than 10% of items with DIF or no items with DIF, they can both be considered invariant with respect to membership in NGG and GG. The only exception was the Self-Direction subscale of the Teacher/Daycare Provider Form, where more than 10% of the total items showed DIF with medium and large effect sizes.

Further confirming the invariance of the instrument, the DTF effect size was negligible in all the subscales mentioned above (.009 < ETSSD < .067), including the Self-Direction subscale of the Teacher/Daycare Provider Form.

According to these results, the TIFs of NGG and GG, computed in the two subscales in which there were DIF items, were almost perfectly overlapping (Figure 2), demonstrating how the functioning of these subscales is essentially the same in the two groups.

Discussion

In the second study, the invariance of the two Teacher Forms of the Italian version of ABAS-3 across teachers who guessed and no-guessed answers was performed. After verifying the IRT assumptions, such as item local independence and the unidimensionality of each subscale of the two teacher forms, DIF analysis applying GRM was performed, even if unidimensionality was not strictly confirmed for all subscales.

In this study, 10 items with DIF were found; in detail, in the Functional Academics, Health and Safety, School Living and Leisure subscales of the Teacher Form and in the Self-Direction, and Health and Safety subscales of the Teacher/Daycare Provider Form there were one or two items (three

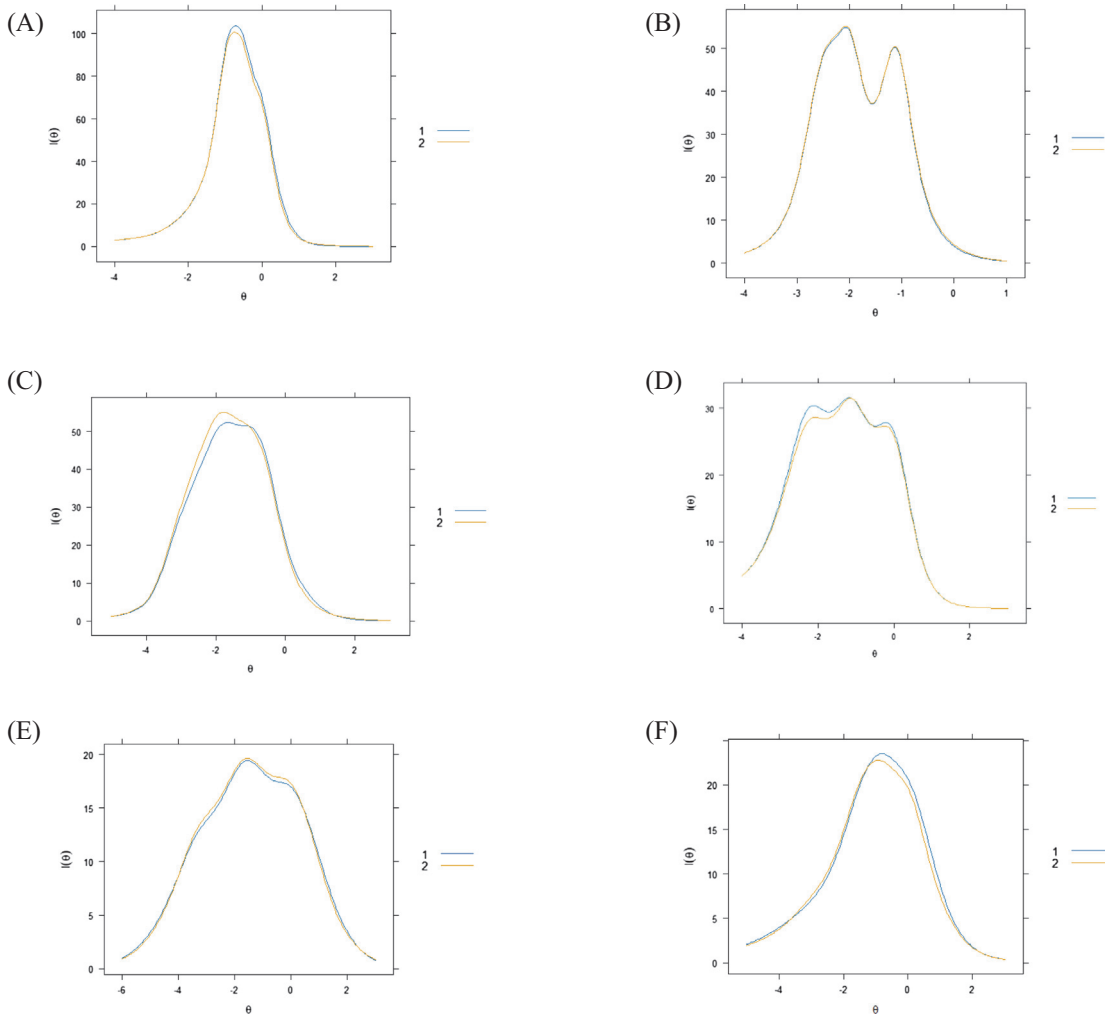

Figure 2. TIFs of NGG (blue line) and GG (orange line) of each subscale of the teacher (A: Functional Academics; B: Health and Safety; C: School Living; D Leisure) and the teacher/daycare provider (E: Self-Direction; F: Health and Safety) forms composed by items with DIF.

only for the Self-Direction subscale of the Teacher/Daycare Provider Form) with a differential functioning across the two groups that differ for the presence of guessed answers. In detail, the presence of biased items observed in the Self-Direction and Health and Safety subscales of the Teacher/Daycare Provider Form aligns with findings from CFA, which highlighted dimensionality issues in these scales. Since DIF can be considered a manifestation of multidimensionality (Teresi, 2006), especially in scales that deviate from strict unidimensionality, the observed DIF in these subscales may result from their inherent multidimensional nature; for that reason, it is possible to suppose that the bias concerning these items is not referred to the guessing issue. Nevertheless, in all cases, items with DIF were less than 10% of the total amount of subscale items, with the only exception of the Self-Direction subscale of the Teacher/Daycare Provider Form, which still showed a negligible DTF effect size. Because a medium to high effect size of DIF was found especially in the Functional Academics and School Living subscales of the Teacher Form and in the Self-Direction and Health and Safety subscales of the Teacher/Daycare Provider Form, the presence of guessed answers in these items should be considered with caution when interpreting the obtained results.

Study 3

Adult Form (ages 16–89 years, 11 months) measures the adaptive functioning of adults in home and community settings, and it can be completed by family members, supervisors, or other respondents who are familiar with the individual in his/her environment (Harrison & Oakland, 2015). The aim of Study 3 was to evaluate the invariance of the Adult Form across NNG and GG groups; for that reason, DIF analyses using IRT in ABAS-3 items of each subscale were performed.

Table 4. CFA fit indices of the two teacher forms of ABAS-3

	Teacher/Daycare provider form			Teacher form		
Subscale	RMSEA	CFI	TLI	RMSEA	CFI	TLI
Communication	.073, CI 90% [0.067, 0.079]	.972	.969	.074, CI 90% [0.069, 0.080]	.960	.954
Community Use	–	–	–	.067, CI 90% [0.059, 0.076]	.983	.980
Functional Academics/Pre-Academics	.055, CI 90% [0.048, 0.062]	.987	.986	.051, CI 90% [0.045, 0.056]	.992	.991
School Living	.086, CI 90% [0.079, 0.093]	.946	.940	.076, CI 90% [0.071, 0.081]	.945	.938
Health and Safety	.122, CI 90% [0.114, 0.130]	.789	.757	.063, CI 90% [0.540, 0.071]	.943	.929
Leisure	.091, CI 90% [0.084, 0.098]	.933	.925	.093, CI 90% [0.085, 0.100]	.913	.894
Self-Care	.092, CI 90% [0.086, 0.098]	.934	.927	.082, CI 90% [0.076, 0.089]	.924	.912
Self-Direction	.148, CI 90% [0.141, 0.154]	.824	.802	.098, CI 90% [0.093, 0.104]	.948	.940
Social	.097, CI 90% [0.091, 0.104]	.938	.931	.097, CI 90% [0.091, 0.102]	.911	.899
Motor/Work	.067, CI 90% [0.060, 0.073]	.965	.961	.001, CI 90% [0.000, 0.088]	.999	.999

Table 5. Items with uniform DIF (*b*) and non-uniform DIF (*a*) according to NGG or GG group membership in the teacher forms of ABAS-3

			Purification phase						DIF analysis							
			*a*DIF			*b*DIF			*a*DIF			*b*DIF			DIF effect size	DTF effect size
Form	Subscale	Item	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*	ESSD	ETSSD
Teacher Form	Functional Academics	22	14.12	1	< .001	–	–	–	23.44	1	< .001	–	–	–	.496	.032
	Communication	10	16.87	1	< .001	–	–	–	.04	1	.001	–	–	–	–	–
	Health and Safety	12	–	–	–	21.59	3	< .001	–	–	–	20.73	3	< .001	.141	.014
	School Living	11	–	–	–	22.72	3	< .001	–	–	–	15.2	3	.002	–	.067
		18	22.76	1	< .001	41.16	3	< .001	2.25	1	.133	21.79	3	< .001	.625	–
		20	–	–	–	31.39	3	< .001				29.16	3	< .001	.275	–
		22	18.94	1	< .001	–	–	–	9.12	1	.003	–	–	–	–	–
	Leisure	6	–	–	–	18.94	3	< .001	–	–	–	17.88	3	< .001	.129	.009
Teacher/Daycare provider form	Self-Direction	8	–	–	–	18.55	3	< .001	–	–	–	18.78	3	< .001	.839	.013
		17	–	–	–	19.49	3	< .001	–	–	–	18.28	3	< .001	.703	–
		23	–	–	–	39.18	3	< .001	–	–	–	34.58	3	< .001	1.019	–
	Health and Safety	10	–	–	–	18.33	3	< .001	–	–	–	18.63	3	< .001	.492	.059
		15	–	–	–	22.25	3	< .001	–	–	–	25.51	3	< .001	.964	–

Table 6. CFA fit indices of the adult form of ABAS-3

Subscale	RMSEA	CFI	TLI
Communication	.056, CI 90% [0.053, 0.058]	.981	.979
Community Use	.063, CI 90% [0.060, 0.066]	.973	.970
Functional Academics	.048, CI 90% [0.046, 0.052]	.991	.990
Home Living	.103, CI 90% [0.100, 0.106]	.909	.899
Health and Safety	.050, CI 90% [0.046, 0.054]	.982	.979
Leisure	.080, CI 90% [0.077, 0.083]	.965	.960
Self-Care	.028, CI 90% [0.025, 0.031]	.995	.995
Self-Direction	.053, CI 90% [0.050, 0.056]	.980	.978
Social	.049, CI 90% [0.046, 0.052]	.983	.981
Work	.037, CI 90% [0.032, 0.041]	.990	.989

Materials and Methods

Participants

A total of 1,656 participants (M_{age} = 43.89 years; *SD* = 19.30; male = 44.57%) completed the Adult Form. Based on the number of guessed answers given, participants were divided into NGG and GG using the previous criteria.

Procedure

The procedure followed is the same as in previous studies.

Results

Results attested that the IRT assumptions were met. In detail, local independence in each subscale was attested

Table 7. Items with uniform DIF (*b*) and non-uniform DIF (*a*) according to NGG or GG group membership in the adult form of ABAS-3

		Purification phase						DIF analysis							
		*a*DIF			*b*DIF			*a*DIF			*b*DIF			DIF effect size	DTF effect size
Subscale	Item	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*	ESSD	ETSSD
Self-Direction	9	–	–	–	18.96	3	< .001	–	–	–	9.29	3	.026	–	–
	15	14.97	1	< .001	–	–	–	2.98	1	.084	–	–	–	–	–
	16	–	–	–	19.13	3	< .001	–	–	–	17.6	3	.001	–	–
	25	–	–	–	17.91	3	< .001	–	–	–	7.41	3	.060	–	–
Functional Academics	12	23.83	1	< .001	–	–	–	14.58	1	< .001	–	–	–	.203	.002
	16	14.97	1	< .001	24.75	3	< .001	.70	1	.403	9.08	3	.028	–	–
	21	–	–	–	32.96	3	< .001	–	–	–	30.83	3	< .001	.065	–
	24	–	–	–	21.59	3	< .001	–	–	–	18.96	3	< .001	.057	–
Communication	5	–	–	–	21.34	3	< .001	–	–	–	24.91	3	< .001	.431	.037
	10	13.12	1	< .001	43.52	3	< .001	.71	1	.401	32.72	3	< .001	.846	–
	13	–	–	–	18.28	3	< .001	–	–	–	15.31	3	.002	–	–
Community Use	18	15.03	1	< .001	–	–	–	2.24	1	.135	–	–	–	–	.003
	22	–	–	–	19.78	3	< .001	–	–	–	19.79	3	< .001	.125	–
	23	–	–	–	18.47	3	< .001	–	–	–	19.84	3	< .001	.043	–
	24	–	–	–	27.47	3	< .001	–	–	–	26.52	3	< .001	.063	–
Self-Care	10	16.86	1	< .001	17.95	3	< .001	5.68	1	.017	6.77	3	.080	–	–
Leisure	8	12.31	1	< .001	–	–	–	6.90	1	.009	–	–	–	–	–
	17	16.72	1	< .001	18.31	3	< .001	1.78	1	.182	3.67	3	.299	–	–
Home Living	9	12.74	1	< .001	–	–	–	9.32	1	.002	–	–	–	–	.040
	11	13.55	1	< .001	–	–	–	21.56	1	< .001	–	–	–	.114	–
	20	–	–	–	33.27	3	< .001	–	–	–	27.68	3	< .001	.111	–
	22	15.28	1	< .001	–	–	–	10.35	1	.001	–	–	–	–	–
	23	12.26	1	< .001	26.75	3	< .001	2.16	1	.142	14.75	3	.002	–	–
	24	21.77	1	< .001	37.27	3	< .001	8.22	1	.004	21.6	3	< .001	.603	–

since no LD value was found to be greater than 10. As shown in Table 6, all models showed a good fit except for the Home Living subscale, where the RMSEA value was greater than .08. Nevertheless, factor-loadings were all adequate (> .30) attesting to the unidimensionality of the subscales: the Home Living has been considered as unidimensional to check the presence of biased items.

Differently from previous studies, few item (Communication: items 10 and 16; Home Living: items 8, 18, 19, and 23; Health and Safety: item 7; Leisure: item 20) showed a significant S-χ^2 value, indicating a poor fit; despite that, the associated effect size indices were acceptable: since RMSEA values of less than .05 indicate that an item adequately fits the GRM (Svicher et al., 2022), considering this index as well, items 10 (RMSEA = .024) and 16 (RMSEA = .025) of the Communication subscale, items 8 (RMSEA = .021), 18 (RMSEA = .023), 19 (RMSEA = .024) and 23 (RMSEA = .019) of the Home Living subscale, item 7 (RMSEA = .024) of the Health and Safety subscale and item 20 (RMSEA = .019) of the Leisure subscale still had an adequate fit. For that reason, these items have been included in the DIF analysis.

Results of the purification phase found some potentially biased items (Table 7): specifically, in the Self-Direction (item 15), Functional Academics (items 12 and 16), Communication (item 10), Community Use (item 18), Self-Care (item 10), Leisure (items 8 and 17) and Home Living (items 9, 11, 22, 23, and 24) subscales non-uniform DIF was found. Moreover, in the Self-Direction (items 9, 16, and 25), Functional Academics (items 16, 21, and 24), Communication (items 5, 10, and 13), Community Use (items 22, 23, and 24), Self-Care (item 10), Leisure (item 17), and Home Living (items 20, 23, and 24) subscales uniform DIF was found. In the second step, all other items with no DIF were used as anchor items, and DIF analysis was carried out. After the iterative process of DIF analysis, only item 12 of the Functional Academics subscale and item 11 of the Home Living subscale continued to show non-uniform DIF. Moreover, the Functional Academics (items 21 and 24), Communication (items 5 and 10), Community Use (items 22, 23, and 24), and Home Living (items 20 and 24) subscales continued to show uniform DIF (Table 7).

The DIF effect size of item 12 of the Functional Academics subscale, and items 11 and 22 of the Community Use subscales were small. Moreover, the DIF effect sizes of items 21 and 24 of the Functional Academics subscale and items 23 and 24 of the Community Use subscale were negligible. Finally, items 5 and 10 of the Communication

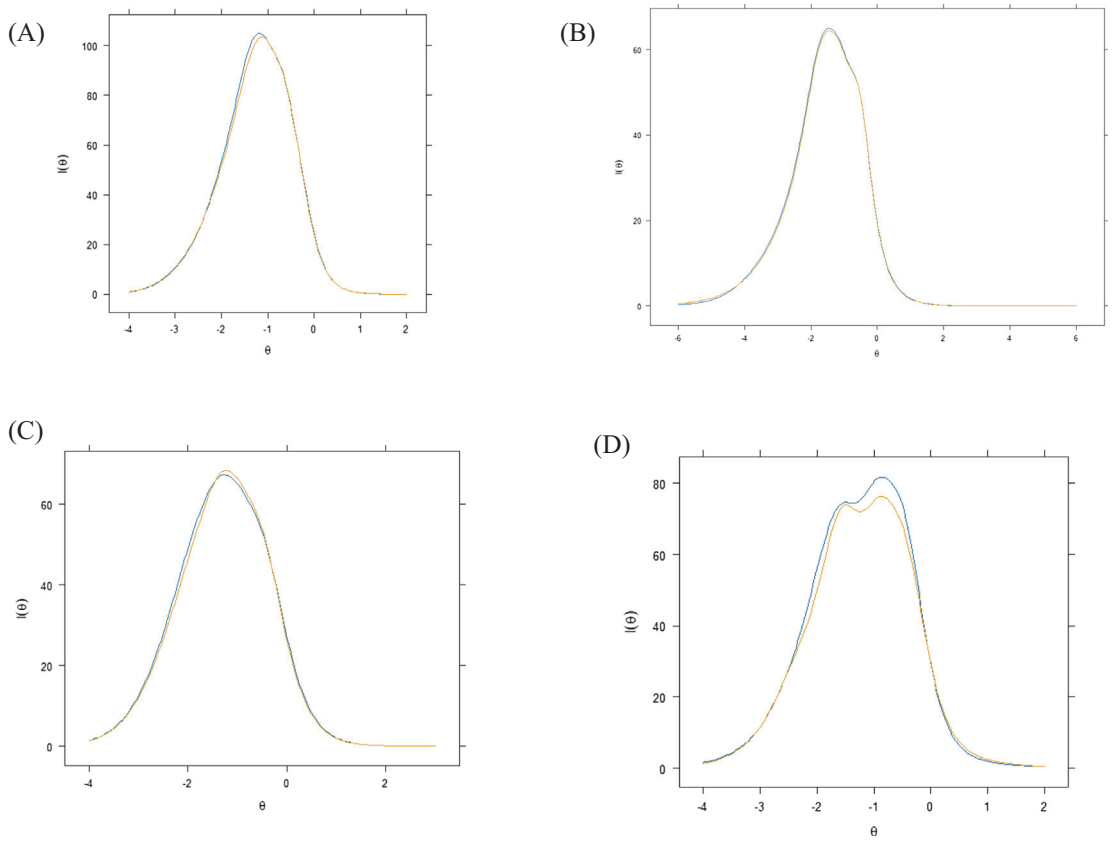

Figure 3. TIFs of NGG (blue line) and GG (orange line) of each subscale with items with DIF of the adult form composed by items with DIF. (A) Functional Academics; (B) Communication; (C) Community Use; (D) Home/School Living.

subscale showed, respectively, medium and high DIF effect sizes, whereas item 24 of the Home Living subscale showed a medium to high DIF effect size. Therefore, only the Community Use scale had more than 10% biased items; despite that, due to the small value of associated effect sizes and the negligible DTF effects size found in all subscales ($.002 < ETSSD < .040$), it can be considered as invariant too with respect to membership in NGG and GG.

According to these results, the TIFs of NGG and GG, computed in the subscales with DIF items, were almost perfectly overlapping (Figure 3). This demonstrates how the functioning of these subscales is essentially the same in the two groups, with the minor exception of the Home/School Living subscale, in which the two curves showed a little difference in the middle.

Discussion

In the third study, the invariance of the Adult Form of the Italian version of ABAS-3 across adults who guessed and no-guessed answers was performed.

After verifying the IRT assumptions, such as item local independence and the unidimensionality of each subscale, DIF analysis applying GRM was performed.

In this study, 11 items with DIF were found; in detail, in Functional Academics, Communication, and Home Living subscales, there were items with a differential functioning across the two subgroups that differed for the presence of guessed answers. In these cases, items with DIF were less than 10% of the total amount of subscale items, or the DIF effect size was negligible. In particular, since especially in the Communication and Home Living subscales, medium to high DIF effect sizes were found, the presence of guessed answers in items with DIF in these two subscales should be considered with caution when interpreting the obtained results. Moreover, three DIF items have been found in the Community Use scale; in this case, the number of biased items was higher than 10%; nevertheless, both items and test effect sizes were small according to a non-strict invariance of the scale. Therefore, it is important to qualitatively evaluate the presence of guessed answers in items with DIF in order to better interpret the obtained results.

General Discussion

Given the wide use of ABAS-3 to assess adaptive behavior (Mecca et al., 2022) and the number of uncertain profiles due to guessed answers, according to the authors' suggestions (Harrison & Oakland, 2015) that indicates four as the maximum number of guessed answers acceptable to consider the profile as certainly valid, the purpose of this study was to assess the test invariance between protocols with guessed answers and those without in the Italian version of ABAS-3.

In detail, DIF analyses between groups of participants who had given no guessed answer – no-guessing group (NGG) – and a group of participants who had given at least one guessed answer – guessing group (GG) were performed. In that way, it was possible to empirically investigate the number of biased items that make the profile invalid, starting from the hypothesis that this cutoff may vary across versions and scales of ABAS-3.

The results showed that all the subscales of ABAS-3 were invariant concerning membership in NGG and GG, indicating that ABAS-3 can be considered an invariant measure of adaptive behavior with respect to the presence of guessing answers. Indeed, only a few items were identified as biased items, and all of them had a negligible or small DIF effect size. Moreover, as a further demonstration of the invariance of ABAS-3 with respect to the presence of guessing answers, in the subscales that showed items with DIF, DTF effect sizes were negligible and the TIFs of NGG and GG were almost perfectly overlapping, indicating that the functioning of these subscales is essentially the same regardless the number of guessed answers.

Despite that, the Teacher Forms and the Adult Form contained the most items with DIF. In most cases, teachers are in contact with their students for only a few hours during the day, so they often don't know or haven't had the opportunity to observe their students' behavior outside the school setting (De Los Reyes et al., 2015). In addition, the results on the Teacher/Daycare Provider Form underscore the importance of considering multidimensionality in the interpretation of DIF results (Teresi, 2006), especially in non-unidimensional scales, to ensure accurate assessment procedure and results interpretation. Regarding the Adult Form, adults often do not share every aspect of their lives with those close to them, and this may result in partial, or no knowledge of some behaviors enacted by the person assessed in certain contexts on the part of the respondent (Connelly & Ones, 2010). For that reason, as recommended by several authors (e.g., De Los Reyes et al., 2012; Hunsley & Mash, 2007), it is important to stress that the respondents may include relatives, partners, roommates, close friends, caregivers, and others in the adult's home because of the respondents need to be familiar with the adaptive skills of the adult being rated to reduce the amount of guessing answers. Indeed, the use and interpretation of reports from multiple informants who are adequately familiar with the person being evaluated are key components of evidence-based evaluation best practices (Dirks et al., 2012; Hunsley & Mash, 2007).

Paying special attention to respondent selection based on these criteria would further improve evaluation, resulting in a positive impact on intervention and treatment design (Makol et al., 2020).

Finally, Parent Forms showed a very small number of items with DIF, probably because parents, especially in the case of children and adolescents, represent reference figures who are very present in their children's lives. Their role allows them to have frequent and constant observation of their children's behaviors in different areas of their lives, consequently enabling them to make an accurate and reliable assessment with respect to their behaviors (Esbensen et al., 2018).

The present study makes an important contribution to reducing the number of ABAS-3 protocols considered invalid. Shedding light on the impact of guessed items on the validity of the test would allow practitioners to avoid treating them as invalid administrations that could be considered valid based on empirical evidence. Because this study indicates that the functioning of ABAS-3 subscales is essentially the same regardless of the number of guessed answers, this study provides evidence to support the validity of the test even in the presence of more than four guessed items: these profiles can be always interpreted to have a valid and reliable measure of adaptive behavior.

Greater clarity on the validity of protocols with the presence of guessed items is particularly useful when ABAS-3 is administered online, a remote administration mode increasingly used by professionals. Indeed, unlike in-person administration, remote administration makes it more difficult for the administrator to ask questions to the test taker to reduce the number of guessed answers (Luxton et al., 2014).

Despite that, another finding of this study concerns the quality of guessed answers. Because of few items showed bias between NGG and NG group, the suggestion is to evaluate which are the guessed answers: in the case of guessed answers in biased items, it is important to evaluate the impact of this answer on the whole profile of the test-taker (Li & Zumbo, 2009). On the other hand, if there are some guessed answers in a no-DIF item, this evaluation is less relevant.

These findings are also clinically relevant, as changes in adaptive behavior ratings are often used to justify clinical improvement or decline, and these kinds of changes can have significant weight in disability assessment processes, criminal court proceedings, and school placement decisions (von Buttlar et al., 2021). ABAS-3 results can be applied in

many ways: assisting in the diagnosis and classification of various developmental and behavioral disorders, identifying functional limitations displayed by children and adults with a variety of challenges or disorders (e.g., Autism Spectrum Disorder [ASD] and Attention-Deficit/Hyperactivity Disorder [ADHD]), documenting a person's eligibility for services and programs (e.g., special education), planning and monitoring interventions designed to improve an individual's adaptive skills and daily functioning, and facilitating research efforts such as program evaluation and treatment outcome studies (Harrison & Oakland, 2015).

Future studies may confirm, extend, or go beyond the limits of the present study. First, in some subscales, IRT assumptions (i.e., the factor structure of Health and Safety and Self-Direction scales of the Teacher/Daycare Provider Form) were not completely fulfilled. This limit is probably related to some groups consisting of a small number of subjects. As a consequence, further studies with larger samples capable of ensuring appropriate response patterns are needed. A second limitation – related to the first one – concerns the non-possibility of dividing the sample into the normative sample and clinical sample due to the too small number of subjects in some groups. Although one would expect to find the same results between the two samples, especially considering that using IRT, there should be no difference in results based on the sample, population-specific studies of clinical samples are needed to empirically confirm this hypothesis.

A third limitation – related to the first one – concerns the DIF items in the Health and Safety and Self-Direction scales of the Teacher/Daycare Provider Form: in these scales there are few biased items and this result may depend on the no-unidimensional factor structure of the scale; therefore, further studies are needed to confirm the monodimensionality of these two scales or to evaluate the presence of DIF items using a multidimensional approach.

In the present study, IRT was applied to determine whether the Italian version of ABAS-3 is an invariant measure of adaptive behavior with respect to the presence of guessing answers. The results of the DIF analysis supported the measurement invariance with respect to NGG or GG membership and, consequently, the validity of ABAS-3 protocols with more than four guessed items in each skill area, even if an inspection of the guessed answers (i.e., biased or no-biased items) is recommended. Therefore, it is reasonable to consider valid all ABAS-3 profiles regardless of the amount of guessed answers; however, professionals should pay attention to which answers are guessed: in the presence of guessing answers to biased items, professionals should carefully investigate these questions with the test-takers. These findings will allow practitioners not to discard valid protocols and make more accurate assessments in the clinical setting, especially when ABAS-3 is administered online.

References

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.). https://doi.org/10.1176/appi.books.9780890425787

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19*(4), 309–321. https://doi.org/10.1177/014662169501900401

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289. https://doi.org/10.3102/10769986022003265

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

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092–1122. https://doi.org/10.1037/a0021212

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin, 141*(4), 858–900. https://doi.org/10.1037/a0038498

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. (2012). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology, 9*(1), 123–149. https://doi.org/10.1146/annurev-clinpsy-050212-185617

Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Embracing not erasing contextual variability in children's behavior–theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 53*(5), 558–574. https://doi.org/10.1111/j.1469-7610.2012.02537.x

Emam, M. M., Al-Sulaimani, H., Omara, E., & Al-Nabhany, R. (2020). Assessment of adaptive behaviour in children with intellectual disability in Oman: An examination of ABAS-3 factor structure and validation in the Arab context. *International Journal of Developmental Disabilities, 66*(4), 317–326. https://doi.org/10.1080/20473869.2019.1587939

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Esbensen, A. J., Hoffman, E. K., Shaffer, R., Chen, E., Patel, L., & Jacola, L. (2018). Reliability of parent report measures of behavior in children with down syndrome. *Journal of Intellectual Disability Research, 62*(9), 785–797. https://doi.org/10.1111/jir.12533

Ferri, R., Landi, M., Borghetto, M., & Lang, M. (2025). *ABAS-3. Adaptive Behavior Assessment System - Terza Edizione*. Hogrefe Editore.

Greiff, S., & Scherer, R. (2018). Still comparing apples with oranges? *European Journal of Psychological Assessment, 34*(3), 141–144. https://doi.org/10.1027/1015-5759/a000487

Harrison, P., & Oakland, T. (2015). *Adaptive behavior assessment system* (3rd ed.). Western Psychological Services.

Harrison, P., & Oakland, T. (2020). *Adaptive behavior assessment system* (3rd ed.)Hogrefe Psykologisk Forlag.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new

alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* (pp. 1–55)(1). https://doi.org/10.1080/10705519909540118
Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3*(1), 29–51. https://doi.org/10.1146/annurev.clinpsy.3.022806.091419
Kim, S., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*(2), 93–116. https://doi.org/10.1111/j.1745-3984.2007.00029.x
Kline, R. B. (1998). *Principles and practice of structural equation modeling*. Guilford Press.
Kline, R. B. (2010). Promise and pitfalls of structural equation modeling in gifted research. In B. Thompson & R. F. Subotnik (Eds.), *Methodologies for Conducting Research on Giftedness* (pp. 147–169). . American Psychological Association.
Kreemers, B., Maljaars, J., Briers, V., Bulteel, S., Storms, G., Maes, B., & Noens, I. (2020). *ABAS-3 Schaal voor adaptief gedrag* [ABAS-3 Adaptive Behavior Assessment System]. Hogrefe Publishers BV.
Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *DOAJ: Directory of Open Access Journals, 30*(2), 343–370. https://doaj.org/article/9ce3a9d065544c53913edbe3a4f4300d
Luxton, D. D., Pruitt, L. D., & Osenbach, J. E. (2014). Best practices for remote psychological assessment via telehealth technologies. *Professional Psychology Research and Practice, 45*(1), 27–35. https://doi.org/10.1037/a0034547
Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728–743. https://doi.org/10.1037/a0018966
Mecca, T. P., Tafla, T. L., Bueno, F. M. B., Valentini, F., Bassetto, S. A., & Teixeira, M. C. T. V. (2022). Transcultural adaptation of the adaptive behavior assessment system (ABAS-3). *International Journal of Developmental Disabilities, 70*(4), 684–695. https://doi.org/10.1080/20473869.2022.2137953
Makol, B. A., Youngstrom, E. A., Racz, S. J., Qasmieh, N., Glenn, L. E., & De Los Reyes, A. (2020). Integrating multiple informants' reports: How conceptual and measurement models may address long-standing problems in clinical decision-making. *Clinical Psychological Science, 8*(6), 953–970. https://doi.org/10.1177/2167702620924439
Nering, M. L., & Ostini, R. (2011). *Handbook of polytomous item response theory models*. Routledge eBooks. https://doi.org/10.4324/9780203861264
Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(S1), 1–97. https://doi.org/10.1007/bf03372160
Schalock, R. L., Luckasson, R., & Tassé, M. J. (2012). *Intellectual disability: Definition, diagnosis, classification, and systems of supports* (12th ed.). American Association on Intellectual and Developmental Disabilities.
Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors*. Paper presented at the Psychometric Society Annual Meeting, Iowa City, IA.
Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40*(4), 331–352. https://doi.org/10.1111/j.1745-3984.2003.tb01150.x
Svicher, A., Palazzeschi, L., Gori, A., & Di Fabio, A. (2022). The gratitude resentment and appreciation test-revised short (GRAT-RS): A multidimensional item response theory analysis in Italian workers. *International Journal of Environmental Research and Public Health, 19*(24), Article 16786. https://doi.org/10.3390/ijerph192416786
Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care, 44*(11), S152–S170. http://www.jstor.org/stable/41219515
Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1–10. https://doi.org/10.1007/bf02291170
von Buttlar, A. M., Zabel, T. A., Pritchard, A. E., & Cannon, A. (2021). Concordance of the adaptive behavior assessment system, second and third editions. *Journal of Intellectual Disability Research, 65*(3), 283–295. https://doi.org/10.1111/jir.12810
World Health Organization. (2007). *International classification of functioning, disability, and health: Children & youth version: ICF-CY*. World Health Organization.
Zander, E., & Bölte, S. (2020). *Adaptive behavior assessment system – third edition*. Hogrefe Psykologiförlaget.

History
Received August 19, 2024
Revision received March 3, 2025
Accepted March 17, 2025
Published online June 12, 2025
EJPA Section / Category Methodological topics in assessment

Authorship
Sofia Santisi, Formal analysis, funding acquisition, investigation, methodology, resources, software, validation, visualization, writing – original draft; Matteo Ciancaleoni, Conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing – review & editing; Michele Borghetto, Rosa Ferri, Maria Landi, Margherita Lang, Data curation.
All authors approved the final version of the article.

Open Science
Open Analytic Code: The scripts, code, and outputs needed to reproduce all of the reported results are not openly accessible. The data is available on request from the author(s).
Open Data: The information needed to reproduce all of the reported results are not openly accessible.
Open Materials: The information needed to reproduce all of the reported methodology is not openly accessible. The material is available on request from the author(s).
Preregistration of Studies and Analysis Plans: This study was not preregistered.

ORCID
Sofia Santisi
https://orcid.org/0009-0008-7152-385X
Matteo Ciancaleoni
https://orcid.org/0009-0006-3626-5619

Sofia Santisi
Department NEUROFARBA
University of Florence
Via di San Salvi, 12
50135 Florence
Italy
sofia.santisi@virgilio.it