the british psychological society
psychological testing centre

Test Review

AID 3 – Adaptive Intelligence Diagnosticum 3

Test Review of AID 3

Reviewers: Charles Johnson & Joanna Horne

Consultant Editor: Sean Keeley

Senior Editor: Charles Eyre

GENERAL INFORMATION AND DESCRIPTION OF THE INSTRUMENT

Test Name: Adaptive Intelligence Diagnosticum 3 (AID 3) (English Edition, version 1.1)

Date of current review: July 2020

Date of previous review: N/A

Original test name: N/A

Authors of the original test: Klaus D. Kubinger & Stefana Holocher-Ertl (German Edition of AID 3)

Authors of the original test: K. D. Kubinger & E. Wurst (German Edition of AID)

Authors of the local adaptation: Klaus D. Kubinger (English Edition, version 1.1)

Local test distributor/publisher: Hogrefe Verlag GmbH & Co. KG

Publisher of the original version of the test: Hogrefe Verlag

Date of publication of current revision/edition: 2014

Date of publication of adaptation for local use: 2017

Date of publication of original test: 1985 (German Edition of AID)

ISBN: 9781854336835

General description of the instrument

The Adaptive Intelligence Diagnosticum 3 (AID 3) is the English language version of a cognitive ability test battery for children and adolescents measuring both basic and more complex cognitive operations. Originally developed in German, the AID 3 aims to provide a profile interpretation of abilities which can be used for such purposes as identifying (partial) performance weaknesses, specific development disorders or learning disabilities. This is in contrast to similar measures which tend to provide a global measure of IQ. The battery consists of 12 sub-tests, measuring 14 test characteristics and 5 add-on tests measuring 6 test characteristics. The tests are administered individually, one-to-one. There is also a supplemental sheet which provides a qualitative assessment of test taker's attitude to work and working and contact behaviours. The English version is applicable for both UK children and adolescents (aged 6:0 – 15:11) and those being educated in English in German-speaking countries. The tests have been developed using item response theory (IRT) and 10 of the sub-tests and one of the add-on tests have adaptive test forms based on branched testing. The other tests use conventional (fixed item) item administration although 8 of the sub-tests can be administered using conventional administration. There are also short forms of 5 of the sub-tests and one add-on test and parallel forms of 8 of the sub-tests. The content and format of many of the sub-tests and add-on tests will be familiar to users of the WISC to which their development was partially thematically related. However, there are other sub-tests and add-on tests which are not related to the WISC and the battery has the added advantage that its adaptive administration allows for much quicker administration (which is approximately 40-75 minutes for the core sub-tests).

Classification

Content domains:

Ability – General

Ability – Manual skills/dexterity

Ability – Learning/memory

Ability – Non-verbal/abstract/inductive

Ability – Numerical

Ability – Spatial/visual

Ability – Verbal

Intended or main area(s) of use:

Clinical

Advice, guidance and career choice

Educational

Description of the populations for which the test is intended

Children and adolescents aged from 6 years 0 months to 15 years 11 months (though can be used with older adolescents in some circumstances)

Number of scales and brief description of the variables) measured by the instrument

12 subtests and 5 supplementary (add-on) tests (see below)

Response mode

- Oral interview
- Paper & pencil
- Manual (physical) operations
- Direct observation

Demands on the test taker:

Manual capabilities

- necessary information given

Handedness

- Irrelevant / not necessary

Vision

- necessary information given

Hearing

- necessary information given

Command of test language

- necessary information given

Reading

- Irrelevant / not necessary

Writing

- Irrelevant / not necessary

Items format

- Open

Ipsativity:

- Not relevant

Total number of test items and number of items per scale or subtest

There are 885 items plus 8 practice items within the entire test. However, as some of the subtests are adaptive, with test-takers completing certain subsets of items based on their age and performance, some of the items are duplicates, although a test-taker would not be administered the same item more than once. In total, there are 574 unique items in the core subtests and 104 unique items in the supplementary subtests. The maximum number of items a test-taker could complete in the entire test is 340. Short forms are available for seven sub-tests and one add-on test.

Test	Conventional	Standard	Short-form
		Adaptive	Adaptive
Sub-tests:			
1. Everyday knowledge	60	15	10
2. Competence in realism	20	15 or 10	10
3. Applied computation	62	15	10
4. Social and material sequencing	19	6	4
5. Immediately reproducing	48	N/A	N/A
6. Producing synonyms	60	15	10
7. Coding and associating	104	N/A	N/A
8. Anticipating and combining	12	6	N/A
9. Verbal Abstraction	61	15	10
10. Analysing and synthesising	25	6	N/A
11. Social understanding and material reflection	61	15	10
12. Formal sequencing	30	9	N/A
Supplementary (add-on) tests:			
5a. Immediately reproducing	14	N/A	N/A
5b. Storing by repetition	18	N/A	N/A
5c. Learning and long-range memorising	1	N/A	N/A
6a. Producing antonyms	60	15	10
10a. Recognition of structures	11	N/A	N/A

Intended mode of use:

Managed mode: Where there is a high level of human supervision and control over the test-taking environment. In CBT testing this is normally achieved by the use of dedicated testing centres, where there is a high level of control over access, security, the qualification of test administration staff and the quality and technical specifications of the test equipment.

Administration mode(s):

- Interactive individual administration

Time required for administering the instrument

Preparation:

Involves both selection of the item blocks for adaptive testing and setting up the materials. This will vary from sub-test to sub-test but will be in the order of one to three minutes per sub-test.

Administration:

Total administration time for the 12 sub-tests with standard adaptive testing where applicable is said to be between 40 and 75 minutes. However, as half the tests are not time limited, it could take longer. Administration times for the five add-on tests are said to vary from 2 to 10 minutes (only one is time-limited) and one add-on test, Learning and Long-Range Memorising, has a 20 minute gap between administrations.

Scoring:

Recording of responses on forms is done during administration and basic scoring involves simple summation of small numbers of items. Total scoring time is likely to be 5 minutes or less.

Analysis:

Scores are transformed by reading values from standardisation tables. The time taken for analysis for each test is probably around 1 minute, approximately 12 minutes for all 12 sub-tests plus approximately 1 minute for each add-on test.

Feedback:

Feedback can be either spoken or written depending on the purpose of testing. The time taken could be anywhere from 15 minutes to 2 hours depending on the nature and complexity of the required report. There are no automated reports.

Indicate whether different forms of the instrument are available and which form(s) is (are) subject of this review

Most of the tests have more than one form (these are not different forms, but parallel forms) although two of the sub-tests and four of the add-on tests are only administered conventionally and so have only one form. All the other tests have a standard adaptive administration form. In addition, seven of the subtests and one add-on test have short-form adaptive administration forms and eight of the sub-tests (Everyday Knowledge; Applied Computation; Social and Material Sequencing; Producing Synonyms; Verbal Abstraction; Analysing and Synthesising; Social Understanding and Material Reflection; Formal Sequencing) have parallel forms of the standard adaptive administration form. Furthermore, because the tests are designed for adaptive administration, different subjects will be exposed to different item sets depending on age group and test performance.

Measurement and scoring

Scoring procedure for the test:

Simple manual scoring key – clerical skills only required

Scores:

For eleven of the 12 sub-tests, scores are the sum of the number of correct responses which are then transformed into 'ability parameters' by reference to a standardisation table. Different standardisation tables are provided based on the age of the test taker. These 'ability parameters' are then converted to T-scores by reference to another set of standardisation tables. Different standardisation tables are provided by gender (male and female) for the Applied Computation sub-test.

In the Immediately Reproducing sub-test, four scores are recorded: the longest correct sequence reproduced forwards; the number of attempts taken to achieve this longest sequence; the longest correct sequence reproduced backwards; the number of attempts taken to achieve this longest sequence. These scores are then transformed into T-scores separately for forward and backwards reproduction.

Three of the add-on tests are also scored by summing correct responses and transforming these to T-scores by reference to standardisation tables. The Learning and Long-Range Memorising add-on test also requires summing of correct responses but also requires counting the number of unsuccessful attempts and calculating difference scores between first and final item administrations. It is the unsuccessful attempts and difference scores which are entered in the standardisation tables. The Recognition of Structures add-on test uses the two-step standardisation process, transforming the sum of scores firstly into 'ability parameters' and then converting these into T-scores.

An overall estimate of cognitive ability can be obtained by examining the lowest and highest T-score and calculating the range of these scores for each test taker. The second-lowest T-score is also considered. These T-scores are all transformed into percentiles using another set of standardisation tables.

Scales used:

- Centiles
- Other (please describe): IRT (Rasch) derived ability parameters
- T-scores

Score transformation for standard scores:

- Normalised – standard scores obtained by use of normalisation look-up table

Computer- Generated Reports

Are computer generated reports available with the instrument?

No

Supply Conditions and Costs

Documentation provided by the distributor as part of the test package:

- User Manual
- Technical (psychometric) manual

Methods of Publication

- Paper

Start – up costs:

The extensive set of materials costs £1076 + VAT and includes:

- Test manuals
- Test administration equipment
- Stimulus booklets
- Picture Cards
- Pattern Sheet
- Puzzle
- Cubes
- Pads
- Picture Board
- 10 recording sheets plus work sheets
- Test Sheet templates, and
- A carrying case

The full set can initially be used with 10 test-takers, with additional recording and work sheets required for further administrations.

Recurrent costs:

Recording sheets (in packs of 25) plus worksheets and test sheets cost £32 + VAT per pack. This works out at £1.28 + VAT for each individual administration.

Prices for reports generated by user installed software: N/A

Prices for reports generated by postal/fax bureau service: N/A

Prices for reports by internet service: N/A

Prices for other bureau services: correcting or developing automatic reports: N/A

Test – related qualifications required by the supplier of the test:

Other (specify): Evidence of competence in the use of psychological tests

Professional qualifications required for use of the instrument:

Practitioner psychologist with qualification in the relevant area of application

Other: Certified training and experience in a relevant discipline; Membership of a professional organisation appropriate to the focus of the test (Health and Care Professions Council, the General Medical Council, the Royal College of Psychiatrists, or the Nursing and Midwifery Council).

EVALUATION OF THE INSTRUMENT

Key to symbols:

[n/a]	This attribute is not applicable to this instrument
0	Not possible to rate as no, or insufficient information is provided
★	Inadequate
★★	Adequate
★★★	Good
★★★★	Excellent

Quality of the explanation of the rationale, the presentation and the information provided

Quality of the explanation of the rationale

Overall rating of the quality of the explanation of the rationale ★★

Theoretical foundation of the constructs	★★
Test development (and/or translation or adaption) procedure	★★
Thoroughness of the item analyses and item analysis model	★★★
Presentation of content validity	★★
Summary of relevant research	★★

Adequacy of documentation available to the user (user and technical manuals, norm supplements, etc.)

Overall adequacy of documentation available to the user (user and technical manuals, norm supplements, etc.) ★★

Rationale	★★
Development	★★★
Development of the test through translation/adaption	★★
Standardisation	★★★
Norms	★★★
Reliability	★★★
Construct validity	★★★
Criterion validity	★★
Computer generated reports	N/A

Quality of the procedural instructions provided for the user

Overall adequacy ★★

For test administration	★★★
For test scoring	★★★
For norming	★★★
For interpretation and reporting	★★★
For providing feedback and debriefing test takers and others	★★
For providing good practice issues on fairness and bias	★★
Restrictions on use	★★
Software and technical support	N/A
References and supporting material	★★
Quality of the procedural instructions provided for the user	★★

Reviewer's comments on the documentation

AID 3 partly follows the Wechsler test concepts, although it is reported that its factor structure does not reflect any pertinent intelligence theory. Whilst intelligence tests traditionally aim to produce an Intelligence Quotient, AID 3 aims to measure the many abilities responsible for intelligent behaviour, with interpretation based on considering the strengths and weaknesses within the entire profile. Rather than a compensation model of intelligence (whereby an IQ score masks deficits in one area that are compensated by strengths in other areas), AID 3 conceptualises intelligence as a deficit model, with the subtest with the lowest score being identified as the most relevant for intervention. There are references to some supporting research, primarily references to the test's author. The manual lacks detail concerning the rationale and development of the tests themselves and the language used is often dense and difficult to follow. It isn't clear from the manual how the different abilities represented by the subtests were determined for inclusion and how the content of these subtests was developed. These subtests are largely similar to the constructs and the tests found in other well-known test batteries such as the WISC and the British Ability scales but there is no coherent explanation of why this particular set of tests was chosen.

The test items were developed for the original German version of the AID in 1985, with minor updates in 2009 and a more thorough revision in 2014. Content validity was checked for the original German AID and revised German AID 3, through practising psychologists judging the practical worth of each specific item, although it's not clear if these experts made any judgement on the overall outline of the test. The manual evidences how the items were developed and selected for one of the original German subtests. For the English version of AID 3 some items were immediately discarded due to cultural specificity and new ones created for item calibration. Rasch model calibration was conducted utilising Andersen's Likelihood-ratio test and graphical modelling, which is clearly outlined in the manual. Items were deleted based on differential item functioning, to produce the required number of items for each subtest. Extra information may be required about the number of items trialled, the trial samples or the initial selection of items for trial. Such details may be contained in the references given in the appropriate sections, but these references are not easy for UK users to access. A further potential difficulty for many UK users is that the discussion of much of the statistical analysis is unclear. Although considerable effort is spent justifying the approaches taken, very little is spent explaining what the statistical analyses actually show or how to interpret them.

The manual provides the total number in the norm group for the English version of AID 3. However, the entire sample did not complete every subtest and the manual does not give the sample size for each subtest. The manual gives some indication that the standardisation involved cluster sampling, with specific schools being determined within 'accessible regions'. However, it is not clear how the regions, or the schools within them were determined. Representativeness of the final sample cannot be established as data regarding SEND, EHCP, FSM and language of the sample are not provided, so cannot be compared to national school parameters. The manual shows clearly how the standardisation subtest results were adjusted for sampling errors, by smoothing out the age-related means, using logarithmic or quadratic functions as appropriate. Analysis was also carried out to test for significant sex differences for each age band on each subtest. Group differences on the basis of ethnicity, language, SEND or FSM are not reported. Separate norm tables are provided for each 12-month age-band (and separately for males and females where appropriate).

The manual provides information on test-retest reliability, parallel form reliability, inter-rater reliability and internal consistency for the German version of AID, but not for the English version. Standard Errors of Estimation are provided for every possible score for each age group on the majority of subtests. Construct validity evidence is provided in the form of exploratory factor analysis (German version of AID), confirmatory factor analysis (German and English versions), discriminant validity (German version of AID), group differences on some subtests (German version of AID). There is no evidence provided for criterion-related validity.

Instructions for administering and scoring the test are clear and detailed. Inevitably, with this being an adaptive test, the administration is more complex than for other similar tests, requiring the items to be scored as they are answered in order to determine the next block of items to be administered, resulting in some degree of page-flicking of the manual being required. Norming instructions are clear, with the raw score (termed as PS – 'Sum of Points') for each subtest being converted into an 'ability parameter' and then transformed into T-scores provided for 12-month age bands. There is some room for human error in this process as the first stage requires correct selection of the student's age, block combination and PS in order to determine the ability parameter; and then the second stage involves selection of the correct age (and sex, in some cases) and PS to determine the T-score. Furthermore, not all ability parameters are included in the T-score tables, so some do require calculation. Standard Errors for ability parameters can also be calculated, and the T-score of a subtest can also be converted into a Percentile Rank. There is a computer program (AID_3_Score) which conducts the score conversions automatically, but this does not form part of this review.

In terms of interpretation, the manual provides a sentence for each subtest, stating what a high score indicates, along with two case studies, which primarily comment on the student's performance in relation to the norm group, and pick out particular strengths / weaknesses. The manual provides some discussion of critical differences between subtest scores. Furthermore, the Recording sheet includes a diagram for screening partial performance weaknesses, distinguishing between the subtests involved in perception, retrieval and utilisation, across different modalities (speech, acoustic, visual, tactile-kinaesthetic, motoric) and across different content (sequencing, discrimination, classification, space localisation). This aids the test user's interpretation by highlighting weaknesses in particular domains. Information in the manual regarding debriefing / feedback is limited to praising the student's efforts and soothing any disappointment.

The manual provides evidence of gender bias and provides separate norm tables for males/females where appropriate. No other sources of bias have been tested. Aside from the appropriate age and location of students to be tested, the manual does not provide clear information regarding restrictions on use in terms of disability, although non-verbal instructions are provided for use with hearing-impaired students. The test does not involve any reading / writing, so literacy levels are irrelevant.

A reasonable reference list is provided, although a number of sources are rather outdated.

Quality of the test materials

Quality of the test materials paper-and-pencil tests

General quality of test materials (test booklets, answer sheets, test objects etc)	★★★
Ease with which the test taker can understand the task	★★★
Clarity and comprehensiveness of the instruction (including sample items and practice trials) for the test taker	★★★
Ease with which responses or answers can be made by the test taker	★★★★
Quality of the formulation of the items and clarity of graphical content in the case of non-verbal items.	★★★
Quality of the materials of paper-and-pencil tests	★★★

Quality of the test materials of CBT and WBT

Quality of the design of the software (e.g. robustness in relation to operation when incorrect keys are pressed, internet connections fail etc.)	N/A
Ease with which the test taker can understand the task	N/A
Clarity and comprehensiveness of the instructions (including sample items and practice trials) for the test taker, the operation of the software and how to respond if the test is administered by computer	N/A
Ease with which responses or answers can be made by the test taker	N/A
Quality of the design of the user interface	N/A
Security of the test against unauthorized access to items or to answers	N/A
Quality of the formulation of the items and clarity of graphical content in the case of non-verbal items	N/A
Quality of the materials of CBT and WBT	N/A

Reviewer's comments on quality of the materials

The test materials themselves are of high quality, professionally produced and attractive for users and intended test takers.

The test container may need to be more robust as it was showing possible signs of damage (e.g. damage to the box handle) and doesn't seem sufficient for the weight of the kit. The internal compartmentalisation is a good idea (making it much easier to locate particular materials than in similar tests), although the dividers may break after limited use. A sturdy bag / backpack or a wheeled trolley would be more practical.

The instructions are generally clear for the test taker and the test administration includes some examples / practice items to ensure clarity. Generally, the test tasks are easy to understand and the style of responding straightforward although there are some exceptions where very careful reading of the test administration instructions is required to work out exactly what is intended. In some cases this may be a result of the instructions being translated from the original German. This is also true of some of the test items that use American English rather than UK English (e.g. ladybugs, trash, elevator) and items in Applied Computation which refer to imperial measurements (yards, inches, pounds, pints) alongside metric ones (and a time of 13 o'clock!), which could cause some confusion to students. The response format is very easy for the test taker.

There are also some items that contain terms / graphics that may cause offence to particular groups and would usually be avoided in educational tests within the UK (e.g. pig, dog, Christmas, craftsman) and some that appear a little dated (e.g. rocking horse) or unusual for the UK (e.g. wine press). Furthermore, some of the correct responses on Producing Antonyms are one-dimensional (e.g. right / wrong could also be right / left; order / obey could also be order / chaos; search / find could also be search / hide).

Norms

Is the test norm referenced? Yes

Norm referenced interpretation

Overall Adequacy: ★★★

Appropriateness for local use	★★★
Appropriateness for intended applications	★★★
Sample sizes (classical norming)	★★★
Sample sizes continuous norming	N/A
Procedures used in sample selection	Cluster sampling controlled by quota sampling (the sampling was controlled for age, gender, and rural-urban proportional rate)
Representativeness of the norm sample(s)	★★
Quality of information provided about minority/protected group differences, effects of age, gender etc.	★★★
How old are the normative studies?	★★★★ (2013)
Practice effects	General Information Given

Is the test criterion referenced? No

Reviewer's comments on the norms

Norm referencing and standardisation are two of the most detailed sections in the user manual. Reference is made to several large samples with reasonably well-described characteristics. This information has clearly been updated over time as the tests have gone through various iterations of the German language versions. Norms for English children are based on more recent data with suitably expressed caution about extending the use of the test to other nationalities.

The norm group for the English version of AID 3 is made up of 570 students from 15 schools in England and 528 students from 10 English-speaking International schools in Germany and Austria, tested between 2010 and 2013. It is questionable whether the latter group are a relevant norm-group for UK students. The English schools were located within 8 local authorities, primarily in Southern England. The AID 3 manual states that the standardisation involved cluster sampling, with specific schools being determined within 'accessible regions', and the sample being selected randomly by quota sampling (based on age, sex and rural-urban proportional rate) from classes within those schools. However, it is not clear how the regions, or the schools within them were determined. Representativeness of the final sample cannot be established as data regarding SEND, EHCP, FSM and language of the sample are not provided, so cannot be compared to national school parameters.

Of the total sample, 48% were female and 52% male. Per 12-month age-band, the numbers (including English and International school students) ranged from 59 (13-year-olds) to 186 (8-year-olds). However, the entire sample did not complete every subtest. The manual does not give the sample size for each subtest, but it does state that the smallest sample size was 450 (Storing by Repetition). This sample split across the 10 age-bands, would give approximately 45 students per 12-month age band (and with potentially only half of these based in the UK), which is rather low, even with continuous norming.

The standardisation is based on Rasch methodology and there is considerable information on the quality of the fit between the Rasch model and the adaptive and non-adaptive versions of all the tests. There is also useful information on the extent to which the fit to the Rasch model is or is not affected by various group differences such as age, gender, geographical location and language version. It is worth noting that in the majority of the analyses provided, the results show that the tests and their items conform to the Rasch model but this is not always the case, particularly with respect to the analyses of group differences where the results show poor fit or significant DIF statistics.

The manual shows clearly how the standardisation subtest results were adjusted for sampling errors, by smoothing out the age-related means, using logarithmic or quadratic functions as appropriate. Analysis was also carried out to test for significant sex differences for each age band on each subtest. Group differences on the basis of ethnicity, language, SEND or FSM are not reported.

Separate norm tables are provided for each 12-month age-band (and separately for males and females where appropriate). Such wide age-bands might be sufficient at the older end of the test's age range (e.g. age 9+, although even here six-month age-bands might be more appropriate), but at the younger end of the test's age range (e.g. age 6-8), where there are more developmental changes taking place, six-month, or even three-month age bands would be more relevant. The manual argues that 12-month age bands are sufficient, but this conclusion is based on research from the original German version of AID with only 8-9 year-olds.

Reliability

Overall Adequacy: ★★★

Overall Adequacy	
Data provided about reliability	Reliability coefficients and standard error of estimation are given for a number of different groups (for each scale or subscale)
Internal consistency:	
Sample size	★★★
Kind of coefficients reported	Other: Standard errors of estimation (Rasch model)
Size of coefficients	★★★
Reliability coefficients are reported with samples which…..	…. match the intended test takers
Test related reliability-temporal stability:	
Sample size	★★★
Size of coefficients	★★★
Data provided about test-re-test interval	A range of time intervals are reported re-ranging from same day to 6 years 5 months. Most of the data is for either approximately 4 weeks or for between 1 and 2 years
Reliability coefficients are reported with samples which……	…. match the intended test takers
Equivalence reliability:	
Sample size	★★
Are the assumptions for parallelism met for the different versions of the test for which equivalence reliability is investigated?	★★★

Size of coefficients	★★★
Reliability coefficients are reported with samples which…….(select one)	**…. do not match the intended test takers, but effect on size of coefficients is unclear**
IRT based method:	
Sample size	★★★★
Kind of coefficients reported	Others: Two types of measure are reported for most of the analyses. The first is Anderson's Likelihood ratio-test statistic (LRT). In addition, graphical representations of the fit of item parameters to the IRT model are given.
Size of coefficients (based on the final test length)	0
Inter-rater reliability:	
Sample size	N/A
Kind of coefficients reported	➢ Not applicable
Size of coefficients	N/A
Other methods of reliability estimation: (Split-Half Co-efficients)	
Sample size	★★★★
Results	★★★★

Reviewer's comments on reliability

A large amount of information is presented demonstrating the reliability of the sub-tests and add-on tests. Most of the information shows the quality of the fit between the test items and the Rasch model but information is also presented on more traditional reliability measures such as split-half reliability coefficients, stability coefficients and short form and parallel version correlations.

Some of the stability coefficients presented in the manual are low even at quite short time intervals. Test-retest reliability is reported for the German versions of AID (1985) / AID 2 (1998) with a four-week interval (n=148). The coefficients range from .57 (AID 2 Storing by repetition) to .95 (AID Everyday knowledge), with a median of .80. Stability coefficients are also reported over a longer interval (average 3 years – minimum 1 year) for AID (1997; n=112), ranging from .39 (Competence in realism) to .79 (Applied computation), with a median of .66. A further study (Parfuss, 2009) investigating stability of AID over varying intervals, produced similar results to those stated above.

Parallel forms: Approximately 50 students from the original (1985) German AID standardisation sample completed parallel forms for those subtests where they are available. The parallel forms are non-adaptive versions of the subtest, where the student completes all unique items. Significant form effects were found for four subtests (Everyday knowledge, Competence in realism, Producing synonyms, and Analysing and synthesising). There is no

evidence of equivalence reliability for the English version of the test, or for later versions of the German test.

IRT analysis is primarily based on the Rasch model, except in the case of the Anticipating and combining subtest, where Master's Partial Credit model is applied. Standard Errors of Estimation are provided for every possible score for each age group on the majority of subtests, with these ranging from 0.54 (Producing antonyms) to 2.41 (Recognition of structures) units of the ability parameter. These equate to approximately 4 and 13 T-scores respectively. It is clear that for those scores with an SEest at the higher end of this range, reliability is rather low.

Inter-rater reliability: 154 students in the German AID standardisation sample (published 1985) were tested twice – half by the same administrator on both occasions and half by different administrators on the first and second testing. Therefore, the sample tested for inter-rater reliability is approximately 77, with one third of these tested using parallel forms. Fischer's multiplicative Poisson model was used to separate test administrator effect from training and parallel form effects. Significant test administrator effects are seen on four subtests: Producing synonyms, Verbal abstraction, Analysing and synthesising, and Coding and associating. This effect may be due to either lack of scoring guarantee and/or lack of test administrator independence. However, there is no evidence of inter-rater reliability for the English version of the test, or for later versions of the German test.

The manual also reports split-half coefficients for the German sample (n=1460), with coefficients ranging from .70 (Competence in realism) to .95 (Everyday knowledge, Applied computation, Analysing and synthesising), with a median of .94. Coefficients are given for only nine of the tests and no internal consistency coefficients for the English version of AID 3 are reported. However, it is possible to estimate coefficient alpha from a Rasch analysis using the Test Information Function (TIF) statistic which therefore also gives an overall indication of the overall reliability of the test and it is a shame that this statistic is not provided for any of the tests in the English version.

Validity

Overall Adequacy: ★★

Construct validity:	
Design used	➢ Exploratory Factor Analysis ➢ Confirmatory Factor Analysis ➢ Testing for invariance of structure and differential item functioning across groups ➢ Difference between groups ➢ Correlations with other instruments and performance criteria ➢ IRT methodology
Do the results of (exploratory or confirmatory) factor analysis support the structure of the test?	★★★
Do the items correlate sufficiently well with the (sub) test score?	★★★★ Note: IRT analyses presented show very good fit to scales, not correlations
Is the factor structure invariant across groups and/or is the test free of item-bias (DIF)?	★★★
Are the differences in mean scores between relevant groups as expected?	★★
Median and range of the correlations between the test and tests measuring similar constructs	0
Do the correlations with other instruments show good discriminant validity with respect to constructs and the test is not supposed to measure?	★★
If a Multi-Trait-Method design is used, do the results support the construct validity of the test (does it really measure what it is supposed to measure and not something else)?	0
Other, e.g. IRT-methodology, (quasi-) experimental designs (describe):	★★★
Sample sizes	0
Quality of instruments as criteria or markers	★★
How are old are validity studies?	10 to 30 years
Construct validity – Overall adequacy	★★

Criterion – related validity:	
Type of criterion study or studies	Post-dictive
Sample sizes	0
Quality of criterion measures	0
Strength of the relation between test and criteria	0
Criterion – related validity – overall adequacy	★★
How old are the validity studies	2 to 30 years

Reviewers' comments on validity

The evidence for the validity of AID 3 is highly variable. The combination of detailed results from the Rasch analyses, analyses of the factor structure and evidence of its differential validity all provide reasonable evidence for the tests' construct validity.

Exploratory factor analysis based on the original German edition of AID produced a 4-factor solution, which the authors suggest fits with the information-processing model of intelligence (Roth et al., 1980), although some subtests did not show a strong fit on any of the factors. Confirmatory factor analysis was conducted, using the same data, to test the goodness of fit to this model, versus other pertinent intelligence theories. This showed that the four-factor solution explained the data better than any other intelligence theory, although it only explains 58% of the variance and was not significantly better than Weschler's model. The confirmatory factor analysis also tested another model based on the 'Survey for identifying specific developmental disorders or learning disabilities' and found this to fit the data even better than the four-factor solution. Exploratory factor analysis based on the English version of AID 3 also shows a 4-factor solution, with a very similar factor structure to the original German version.

Evidence for convergent validity is provided in a very patchy fashion through reference to a range of studies where only summaries of the main findings are provided, but no quantified data. The manual explicitly states that no criterion validation has been done but that an alternative approach to validation has been adopted, namely "an evaluation of diagnosis-specific enhancement". Discriminant validity is evidenced by correlations between the German version of AID and several achievement tests / personality measures, with the resulting coefficients suggesting that AID does not effectively measure mechanic-technical comprehension ($r<.35$), visual space ($r<.30$), alertness ($r<.35$), or personality traits (general anxiety, self-consciousness, impulsivity, self-complacency, inferiority; $r<.33$), and only two subtests (Formal sequencing and Analysing and Synthesising) correlate with pertinent matrices tests, as would be expected. It would be recommended that evidence of both convergent and discriminant validity could be more appropriately provided by correlating all the subtests of the English version of AID 3 with all the subtests from similar tests currently used in the UK (e.g. BAS-III and WISC-V-UK) in order to determine the pattern of subtest intercorrelations.

There is rather limited evidence provided of group differences on specific subtests, e.g. between good vs poor discriminators, poor vs adequate special memory, good vs poor space localisation. A large number of references are provided for studies which have examined group differences, and, to a lesser extent, diagnostic-specific enhancement but almost no quantitative information is given about effect sizes or diagnostic accuracy. However, these studies are, generally, rather old and based on the original German version of AID. This information may be contained in the references given but, again, these are difficult to access. From the information provided in the user manual, the user is really being asked to take on trust the effectiveness of the tests in achieving their purpose.

DIF effects were noted during the item analysis stage comparing the German / English samples, which resulted in the removal of some items where significant differences between the two samples were indicated. Rasch model tests on the final version of each subtest, partitioning for sex, age, country and language are presented. Although significant effects are shown on all criteria on all subtests, the authors do argue that graphical model checks do show model confirmation.

There is no evidence provided for criterion-related validity.

Final Evaluation

Evaluative report of the test:

The overall value of the AID 3 test is difficult to determine. The German language versions of the tests have a long history of use and have been extensively researched. There is evidence to suggest that the tests have been carefully designed and are reliable with good construct validity. They may also have good criterion validity, particularly given their similarity to other well-established tests, but it is difficult to be sure of this because of the absence of good quality validation evidence in the user manual and the difficulty of accessing relevant references.

The main advantage of the AID 3 tests over other, similar, test batteries is the adaptive testing provided by their branched testing design. This means that a large number of different tests can be administered in a relatively short time with little loss in accuracy. The adaptive testing design may be an attempt to make the testing time shorter but it is similar to the approximate test times for the WISC-V UK (Wechsler Intelligence Scale for Children – Fifth UK Edition) core subtests or for the those of the BAS-3 (British Ability Scales – Third Edition). A potential testing time of 75 minutes for children at the younger end of the test's age range is rather long, even with rest breaks. The WISC-V also allows for web-based administration, scoring and reporting, which may be preferable to students. Whilst similar tests (e.g. WISC / BAS) have been extensively redesigned over the years, to fit with current research, AID 3 appears to be measuring the same abilities as the original version, albeit with revisions to items. That said, the AID 3 does have a different aim to traditional intelligence measures, specifically to identify particular areas of weakness relevant for intervention, rather than provide a global measure of IQ.

It is questionable whether half the norm group (being English-speaking International schools in Germany and Austria) are relevant for UK students and the representativeness of the final sample is not established. The sample per age band on some subtests is also rather small, and 3- or 6-month age bands would be more appropriate for the younger age groups.

Reliability is generally established for the German version of AID, rather than for the English version of AID 3. Standard Errors of Estimation are reported for the English version, but some are rather large. Apart from confirmatory factor analysis, evidence of construct validity is also primarily based on the German version of AID and there is no evidence of convergent validity or criterion-related validity.

Conclusions:

The AID 3 tests are capable of being a valuable addition to the tools used by educational, clinical and child development psychologists. Those already trained in the use of similar instruments, such as the WISC, will find the adaptive format and its possible reduction in test administration time to be attractive features.

AID 3 is clearly suitable for use in research purposes and may be useful in assessment, by an experienced tester, for the purpose of identifying strengths / weaknesses within a student.

For more widespread use of AID 3 within the UK, the following recommendations would be made:

- That the authors clearly consider the rationale for the inclusion of the specific subtests within AID 3 to ensure that these reflect the abilities responsible for intelligent behaviour, and provide some justification of this in the manual
- That a test developer in the UK be consulted to check the appropriateness of items and advise on the deletion of unsuitable items
- That a full UK-wide standardisation is undertaken with a larger representative sample
- That the reliability and validity of the English version of AID 3 are clearly established
- That evidence of both convergent and discriminant validity are established by correlating all the subtests of the English version of AID 3 with all the subtests from similar tests currently used in the UK (e.g. BAS-3 and WISC-V-UK).

The complexity of the underlying concepts in both the tests' content and in the user manual suggest that only experienced and knowledgeable will be able to use the tests effectively in the way recommended and will need to not only be qualified test users at EFPA level 2 but also be experienced practitioners with a wide knowledge of relevant theory.

Recommendations:

- Only suitable for use by an expert user (exceeding EFPA User Qualification Level 2) under carefully controlled conditions or in very limited areas of application.