Psychological Test and Assessment Modeling, Volume 65, 2023 (2), 195-205

A reasonable approach to check a psychological test's long ago standardization – applied for the Adaptive Intelligence Diagnosticum (AID 3)

Klaus D. Kubinger & Thomas Suster

University of Vienna, Faculty of Psychology

Abstract

Due to the DIN 33430 [Requirements on procedures for the assessment of professional aptitude], a psychological test's standardization must be verified every eight years. However, in particular for tests that are to be individually administrated, standardization means a vast and very expensive undertaking. Therefore, sampling new data only for checking a psychological test's long ago standardization is to be minimized or preferably avoided. This paper suggests to use data of case reports from pertinent institutions instead. It is outlined that this approach is not only economical but also very reliable – although such data is most likely not representative with regard to the population in question: By this means easily not only a single but rather several surveys are possible within the time-line from the original standardization data sampling to the actual point in time; they disclose whether some linear or at least some monotonically decreasing or increasing progression of the mean of test-scores occurs or just random fluctuations of the test-score level take place. Even if the results force the revision of the test's standardization, this approach is probably of further use; regression analysis could lead to a predicted value of the test-score's mean and standard deviation in the next calendar year(s), which can be the basis for a re-standardization. Suitability of the current suggested approach is illustrated through an example that concerns a single subtest of the intelligence test-battery *Adaptive Intelligence Diagnosticum* (AID 3; Kubinger & Holocher-Ertl, 2014).

Keywords: test standardization; representative sampling; DIN 33430; two-way analysis of variance (cross-classification, mixed model); Adaptive Intelligence Diagnosticum (AID 3)

Author Note

Klaus D. Kubinger, PhD., Professorial Research Fellow, University of Vienna, Faculty of Psychology, Liebiggasse 5, 1010 Vienna, Austria. email: klaus.kubinger@univie.ac.at

Introduction

According to the standards DIN 33430 (DIN Deutsches Institut für Normung e.V., 2002, 2016)[1], a psychological test must prove the norming (i. e. "standardization" of test-score norms) is up to date every eight years. Although that number of years (eight) is completely arbitrary, the demand for an early standardization's verification is reasonable, merely with regard to the so-called Flynn effect (cf. Flynn, 1984, 1996; notice, that there are nowadays results which indicate a respective trend just in the opposite direction: Dutton, van der Linden, & Lynn, 2016). As a consequence, test-authors and test-publishers, respectively, are obliged to recurrently check the appropriateness of their test's standardization.

However, this can be very expensive, as standardization of a psychological test typically requires about 2000 representatively sampled testees. Standardization means a vast undertaking, particularly for tests that are provided for individual- but not for group-testing. For this, methods are needed to minimize the effort for checking a psychological test's long ago standardization.

The statistical technique of so-called sequential testing could certainly help: here, data is sampled one after the other until either the null- or the alternative hypothesis is to accept and the other to reject – otherwise additional data have to be sampled (see e. g. Rasch, Kubinger, & Yanagida, 2011). However, this would only save a lot of the required sample size if the null-hypothesis is finally accepted, that is, the validity of the early standardization was approved up until now. Otherwise, in the case that the alternative hypothesis is finally accepted, there is all the more reason for a very large, new sample of testees in order to produce a new, highly accurate, current standardization. Beyond that, it is important to take into account that test-standardization is rather a question of the optimal sampling method in regard to the representativity of certain sub-populations than a question of the statistical approach. Hence, sampling new data just for the sake of checking whether a certain test's standardization is still appropriate – and potentially, as a consequence of this, sampling a large data set in order to establish a new standardization – should be avoided.

Therefore, this paper suggests an alternative approach. It is based on the data of case reports which are given at the test authors' disposal by practitioners of pertinent institutions. As explained in the following and illustrated by empirical data, this approach proves to be economical and most reliable.

[1] which were the basis for the world-wide being in force ISO 10667 (*Assessment service delivery – Procedures and methods to assess people in work and organizational settings*; ISO International Organization for Standardization, 2011a, b)

Method

Using data of case reports from several institutions is most likely not representative for the population in question. This is not only due to the fact that sampling psychological consulting institutions always is arbitrary because of their willingness to cooperate. It is also due to the fact that the clientele of such institutions basically differ from the remaining population. Yet, a lack of representativity in the resulting sample is not harmful when it comes to the given problem: "Is the standardization of a specific test up to date or is it not?"

Obviously, most of the pertinent institutions differ even from the very beginning from the given norms in relation to the mean of the test scores (and maybe also with respect to their standard deviation), due to the indication-based selection of their people. In other words, the analysis should not focus on whether such institutions are able to uphold the test performance level of the population since test publication, but rather whether such institutions uphold their own specific test performance level since then or not. Given there is a universal shift of the population's level, then such institutions should consistently disclose that shift, irrespective of their deficient representativity. This is at least the case for such institutions that did not substantially change their clientele in the meantime.

The indicated approach has some advantages compared to the traditional approach of testing a new (representative, large) sample. While the latter, the traditional approach, is restricted to a single point in time for the determination of the test-score level, which is then to be compared to the original level of the psychological test in question, the former is not. Instead, the indicated approach offers several surveys within the timeline from the original standardization data sampling to the point in time when the examination should be carried out. According to DIN 33430, this amounts to seven to eight units of test-score level determination. As an advantageous consequence, even the course of test-score level-shifts can be analyzed which will likely disclose whether, over time, a linear shift or a monotonically decreasing or increasing progression of the mean of test-scores or just random fluctuations of the test-score level take place. By comparing only two mean test-scores, the traditional approach could however result in a non-recurring, random (but perhaps significant) effect, which could be erroneously interpreted as a relevant effect. As an example, take the coronavirus pandemic in the year 2020 as the point in time to check for a certain test-standardization's appropriateness: Due to multiple lockdowns, psychological consulting was most likely reduced, and this might have been varying between testees with a specific test performance level and those with different levels.

In the ideal situation, applying the indicated approach would result in a data design with three factors: "age of testee", "calendar year of test administration", and "institution". In case several institutions are constrained to specific age-groups of testees, then however the respective number of cells in such a three-dimensional design would disclose a lot of missing data. Hence, it might be suggested that the factor "age of testee" has no effect and is therefore disregarded. Anyway, given the test-scores are

standardized for each age-group separately, all testees, irrespective of their ages, can be pooled. This results in a cross-classification design, which is assigned as a mixed model A × **B** within the terminology of analysis of variance. That is, the factor levels of A ("calendar year of test administration") are fixed, but those of **B** ("institution") are random – i.e. the latter are randomly drawn from a population with a very large number of factor levels (cf. e.g. Rasch, Kubinger, & Yanagida, 2011). Such a design refers to three null-hypotheses, however only one of them is of relevance: "The mean test-scores do not differ with regard to the calendar years of test administration". The second null-hypothesis ("The mean test-scores do not differ with regard to the institution") is most likely to be rejected due to the previously given assessment that such institutions have their own specific test performance level. Finally, the third null-hypothesis ("The mean test-scores do not differ with regard to certain combinations of institution and calendar year") is hardly of interest because: if the first null-hypothesis holds, the additional rejection of the third only proves that the test performance level differs in some institutions to other institutions in certain calendar years to a different degree than in other years, which rather reveals an institution-specific change of the clientele over the years; but if the first null-hypothesis is rejected, then the original test-standardization has already proven to be inappropriate anyway – bear in mind that with regard to the theory of the analysis of variance, the first null-hypothesis refers to general mean differences between calendar years, irrespective of any general mean differences between institutions.

When analyzing data accordingly, some critical issues arise:

1) If there is not only a single test but a test-battery with several subtests, then the analysis in question multiple applied means to take a comparison-wise risk for hypothesis testing into account instead of a study-wise risk. This means, above all, that risking a type-I-error with probability α for each subtest leads to a (theoretically) non-calculable overall type-I-risk. But this does not matter as each subtest's standardization is actually of its own concern.
2) If the null-hypothesis "The mean test-scores do not differ with regard to the calendar years of test administration" has to be rejected, then of course some post-hoc tests could be applied in order to identify the (groups of) calendar years which differ. However, not every grouping of calendar years would make sense with regard to the fundamental null-hypothesis, which is essentially of interest but not quoted so far: "There is no progression of monotonically decreasing or monotonically increasing mean of test-scores with the running calendar year of test administration" – be aware, this hypothesis also includes the case that the respective means disclose a difference solely due to a kind of "jump" between two successive years. That is, the evidence of a significant decreasing mean of the test-scores from one year to the next, accompanied later by a significant increasing mean of the test-scores from one year to the next (or *vice versa*), does not support this fundamental null-hypothesis, but rather indicates some changes in the clientele: Most likely, some (arbitrary) events

happened to occur, which, however, hardly cause permanent effects on the mean of the test-scores. Such fluctuations do not at all justify a new standardization of the test. Hence, at least in addition to the referenced post-hoc tests, analyses that examine the sequence of the mean of test-scores are needed. For this, a graphic may serve as well as a regression analysis for testing the null-hypothesis: "The slope of the (linear) regression coefficient $\beta = 0$".

3) The last null-hypothesis is most illustrative when tested for each institution separately. However, in doing so, a comparison-wise type-I-risk occurs. Hence, either an alpha-adjusting procedure has to be applied or the interpretation of the results *in toto* must consider a certain number of significant results only by chance. Given, for instance, $k = 20$ institutions and a comparison-wise type-I-risk $\alpha = .05$, then the probability for mistakenly rejecting the null-hypothesis at least once amounts to $1 - (1 - .05)^{20} = .6415$, which is in almost two-thirds of such research studies; and for mistakenly rejecting the null-hypothesis at least twice, the probability amounts to .2642, which is in more than a quarter of such research studies.[2] Above all, any conclusion depends on whether or not (significant) non-zero slopes are consistently, for every institution, positive or consistently negative – and on the examination of whether there actually is a progression of a monotonically decreasing or monotonically increasing mean of test-scores with the running calendar year of test administration, but no fluctuation of these means over the years that appears random.
4) It is most convenient to pool the data within each institution, that is explicitly not taking into account any differential progression of the mean of test-scores over the years in particular concerning male and female testees or younger and older ones. For this, the null-hypothesis „The slope of the (linear) regression coefficient does not differ between the mentioned groups" should be tested exemplarily. Weber (1980), for instance, gives the respective, seldom used formula.

Results of an Example

An example serves to illustrate the approach being discussed.

The test in question is the *Adaptive Intelligence Diagnosticum* (AID 3, in the German version 3.1; Kubinger & Holocher-Ertl, 2014). This is an intelligence test-battery, which is to be individually administrated for children and adolescents from the age of 6 up to 16. There are twelve sub- and five add-on-tests, from which only the subtest

[2] this calculation according to: https://matheguru.com/stochastik/binomialverteilung.html

Applied Computation will be considered here. A public calling *via* professional psychologists' associations in Germany and Austria as well as *via* the publisher's and the first author's mailing lists raised data from 14 institutions and 5203 testees altogether – tested between 2014 and 2020.[3] Table 1 shows the number of testees for each calendar year and each (anonymized) institution. Sometimes data was not available from each calendar year.

Table 1:

The number of testees for each calendar year and each (anonymized) institution (all in all $n = 5203$ testees)

	2014	2015	2016	2017	2018	2019	2020
BY	18	46	47	54	58	63	43
BL	20	53	37	54	41	49	43
EA	28	45	42	39	36	38	23
LP	14	38	27	36	35	34	30
TY	-	72	85	81	30	26	106
JD	-	25	55	76	75	86	57
PQ	-	13	12	12	19	22	15
VT	-	22	19	21	45	51	91
TL	-	-	56	121	184	184	160
IP	-	-	17	20	22	22	-
PJ	-	-	-	207	302	420	300
DU	-	-	25	46	40	-	-
UY	-	-	-	-	138	399	165
HM	-	-	-	-	53	61	54

Analysis of the data using the two-way analysis of variance according to the cross-classification design as indicated above (mixed model A × **B**; A the fixed factor "calendar year of test administration" with the seven factor levels 2014, 2015, … 2020,

[3] As a whole, there were (validated) data from 5940 testees, but some of them came from those 18 institutions with a statistically too small number of cases. Also, data from those testees of the above-mentioned 14 institutions had to be deleted if their number of testees in a certain calendar year was too small.

and **B** the random factor "institution" with the 14 factor levels BY, BL, … HM) yielded the results as given in Table 2.[4]

Table 2:

The resulting *p*-values for the AID 3's subtest *Applied Computation* according to the two-way analysis of variance (mixed model A × **B**), above all referring to the null-hypothesis concerning factor A: "The mean test-scores do not differ with regard to the calendar years of test administration"; furthermore, regarding the null-hypothesis concerning the interaction effect A × **B**: "The mean test-scores do not differ with regard to certain combinations of institution and calendar year", as well as the null-hypothesis concerning factor **B**: "The mean test-scores do not differ with regard to the institutions" – the number of testees are distributed over the calendar years 2014 to 2020 as follows: 80, 314, 422, 767, 1078, 1455, 1087; $\alpha = .01$.

	A	A × **B**	(**B**)
Applied Computation	.263	.160	.000

According to Table 2, the mean test-scores in the AID 3's subtest *Applied Computation* do not significantly differ depending on the calendar year of the psychological test-application ($p = .263 > \alpha = .01$). That is, the long-ago test-standardization is still appropriate. – In line with the expectation the institutions' mean test-scores differ significantly; and there is no interaction effect with regard to certain combinations of institution and calendar year.

As, however, already indicated, testing the null-hypothesis "The mean test-scores do not differ with regard to the calendar years of test administration" only takes mean difference(s) between any calendar year into consideration, but not a feasible monotone progression (trend). Although this hypothesis holds in the given example, a regression analysis accompanied by a graphical illustration of the mean test-score sequence over the years seems useful: Perhaps a currently not significant trend would be established, which could be become relevant in some years later.

Figures 1 and 2 now show, exemplarily for both the institutions with the most testees, the mean test-score sequence over the years (and also the sequence of the characters "mean plus …" and "mean minus a standard deviation of the test-scores"). Additionally, the regression line for the mean test-scores and the calendar year is given, as well as the regression coefficient of the respective slope. Table 3 shows the *p*-values of the respective significance test for each institution (comparison-wise $\alpha = .01$).

[4] Calculation with SPSS Version 27 was done by the second author within his Master Thesis which was supervised by the first author.

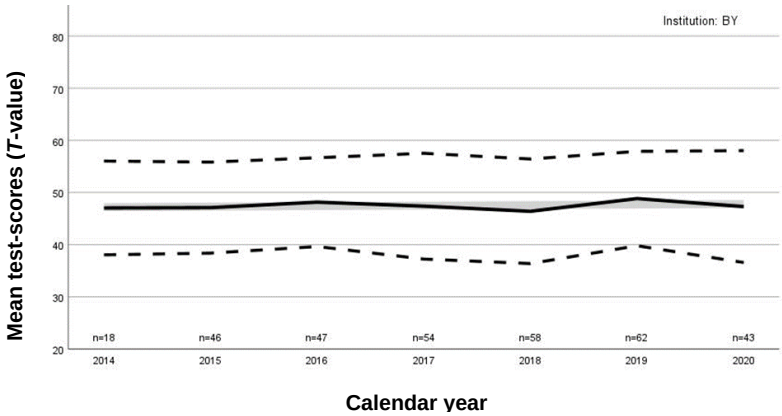

Figure 1

AID 3's subtest *Applied Computation*, Institution BY: Mean test-score sequence over the years (full black line), the sequence of the characters "mean plus …" and "mean minus a standard deviation of the test-scores" (dashed black lines), and the regression line for the mean test-scores and the calendar year (bold light gray line; regressand: mean test-scores [*T*-value], regressor: calendar year; slope $b = 0.10$, $p = .73$; $n = 328$). [Reprint with kind permission from *Hogrefe Verlagsgruppe*]

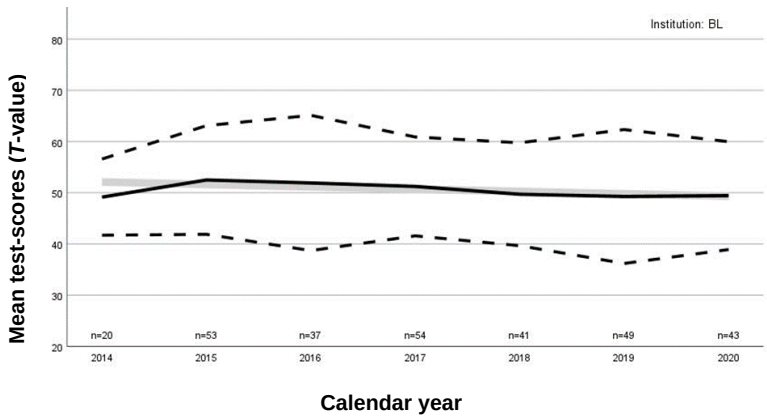

Figure 2

AID 3's subtest *Applied Computation*, Institution BL: Mean test-score sequence over the years (full black line), the sequence of the characters "mean plus …" and "mean minus a standard deviation of the test-scores" (black dashed lines), and the regression line for the mean test-scores and the calendar year (bold light gray line; regressand: mean test-scores [*T*-value], regressor: calendar year; slope $b = -0.46$, $p = .17$; $n = 297$). [Reprint with kind permission from *Hogrefe Verlagsgruppe*]

Table 3:

The resulting *p*-values for AID 3's subtest *Applied Computation* regarding the null-hypothesis "The slope of the regression line for the mean test-scores and the calendar year is zero" for each institution (comparison-wise $\alpha = .01$). If the slope is positive, the result is underlined.

Institution	BY	BL	EA	LP	TY	JD	PQ	VT	TL	IP	PJ	DU	UY	HM
Applied Computation	.73	.17	.02	.72	.01	.71	.90	.17	.33	.13	.11	.28	.30	.72

Table 4 additionally shows the resulting *p*-values for a single institution (PJ) when the slopes between female and male testees and younger and older testees, respectively, are compared ($\alpha = .01$).

Table 4:

The resulting *p*-values for AID 3's subtest *Applied Computation* (institution PJ only) regarding any sex-specific or age-specific slope of the regression line for the mean test-scores and the calendar year (comparison-wise $\alpha = .01$; according to Formula 14.2.7. in Weber, 1980).

	number of testees altogether	female *vs*. male	age of the testee 6-11 *vs*. 12-16
Applied Computation	1121	.05	.63

The already given conclusion according to Table 2, that is, the long-ago test-standardization is still appropriate, can be backed by Figures 1 and 2 and Table 3: A monotone progression (trend) is not visually recognizable for the sequence of the mean test-scores over the years, and the respective regression line's slope hardly differs from zero, at least not significantly. Furthermore, roughly half of the institutions disclose a (non-significant) positive slope and the other half a (non-significant) negative slope, which impressively contradicts any systematic progression. As a consequence, no severe change of the mean test-scores is to be expected in the near future – given, no "big political-societal event" happens. By the way, also the sequence of the characters "mean plus …" and "mean minus a standard deviation of the test-scores" do not disclose any systematic progress. Finally, as Table 4 shows exemplarily, pooling institution's data irrespective of sex or age of the testee is actually justified.

Discussion

The given example illustrates that the presented approach for regularly checking a psychological test-standardization's appropriateness works well. Overall, the principle of sampling data continuously since the time of test publishing is compelling. In contrast to sampling data only once, approximately eight years after test publication, this principle more or less guarantees that a non-recurring random effect (some random fluctuation of the test-score level) does not lead to a mistaken interpretation.

In the given example the test-standardization's appropriateness is confirmed.[5] That is the best case. However, if the results force test-author (and test-publisher) to revise the standardization, then the suggested approach is probably also of use for a re-standardization. This is true if there is actually an unequivocal trend of the test-score level, uniform for all evaluated institutions. Then, applying a (linear or even non-linear) regression analysis leads to a predicted value of the mean test-score of the next calendar year(s). Assessing the regression lines analogously for the characters "mean plus …" and "mean minus a standard deviation of the test-scores" would result in a proper prediction of the standard deviation, too. Given that the test-scores were originally normally distributed (e.g. due to a *T*-score transformation), a re-standardization could easily be undertaken now by using that (predicted) mean and standard deviation.

References

DIN Deutsches Institut für Normung e.V. (2002). *Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen: DIN 33430* [Requirements on precedures for the assessment of professional aptitude]. Berlin: Beuth.

DIN Deutsches Institut für Normung e.V. (2016). *Anforderungen an berufsbezogene Eignungsdiagnostik: DIN 33430* [Requirements on assessment of professional aptitude]. Berlin: Beuth.

Dutton, E., van der Linden, D., & Lynn, R. (2016). The negative Flynn Effect: A systematic literature review. *Intelligence, 59*, 163- 169.

Flynn J. R. (1984) The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1996). What environmental factors affect intelligence: The relevance of IQ gains over time. In D. K. Detterman (ed.), *The environment – Current topics in human intelligence* (pp. 17–29). Westport: Ablex-Publishing.

[5] Readers interested in the results of applying the suggested approach for the other AID 3's subtests are referred to Kubinger (2023, in print).

ISO International Organization for Standardization (2020a). *International Standard ISO 10667-1-2020: Assessment service delivery – Procedures and methods to assess people in work and organizational settings (Part 1: Requirements for the client).* Genf: ISO.

ISO International Organization for Standardization (2020b). *International Standard ISO 10667-2-2020: Assessment service delivery – Procedures and methods to assess people in work and organizational settings (Part 2: Requirements for service providers).* Genf: ISO.

Kubinger, K. D. (2023, in print.) *Manual zum AID 3 (Version 3.2) von K. D. Kubinger & S. Holocher-Benetka* [Manual of the AID 3, version 3.2, by K. D. Kubinger & S. Holocher-Benetka]. Göttingen: Hogrefe.

Kubinger, K. D. & Holocher-Ertl, S. (2014). *Adaptives Intelligenz Diagnostikum - Version 3.1 (AID 3)* [Adaptive Intelligence Diagnosticum]. Göttingen, Germany: Beltz.

Rasch, D., Kubinger, K. D. & Yanagida, T. (2011). *Statistics in Psychology – Using R and SPSS.* Chichester: Wiley.