30 Advanced Statistical Analyses

Miguel A. Padilla

Overview
Statistical models offer much flexibility and many fall under several general umbrellas. The linear mixed model is one such model, and it can be specified to answer vastly different research questions. Three such linear mixed model specifications (or methods) are the hierarchical linear models used when there are clustered data structures; generalizability theory used for evaluating the reliability or consistency of a measurement process; and equivalence testing used for investigating the similarity between conditions. Here, each method is presented in the context of healthcare simulation with a worked-out example to highlight its central concepts while technical details are kept to a minimum.

Practice Points

- A linear mixed model (LMM) is an umbrella model that can be formulated to answer a variety of research questions.
- A LMM formulated to account for clustered (nested) data structures is called a hierarchical linear model.
- Generalizability theory is a form of a LMM formulated to evaluate the consistency of measurement (assessment).
- Equivalence testing is another form of a LMM formulated to measure the equivalence of groups.
- Any statistical package with a LMM routine can obtain the basic results for the three methods discussed.

M. A. Padilla (✉)
Department of Psychology, Old Dominion University,
Norfolk, VA, USA
e-mail: mapadill@odu.edu

Introduction

Healthcare simulation is a rapidly growing field due to advancements in technology and research methodology. Statistical methods are a major part of research methodology, and three advanced statistical methods are presented here. Statistical models have been developing for over a century, and many of them can fall under several general umbrellas. The linear mixed model (LMM; or just mixed model) is one such umbrella model [1]. It is called a LMM because it can model any combination of random and fixed effects. A random effect is when the levels of a variable (or factor) can be thought of as being sampled from a corresponding population. For example, if data are collected from different medical centers and "center" is in the model, then "center" can be thought of as a random effect. By contrast, an effect is called fixed if the levels in the study represent all the possible levels of a variable (or factor). Some examples of fixed effects include gender (male, female) and treatment method (treatment, placebo). The modeling flexibility of LMMs allows them to be specified to answer a variety of research questions. LMMs have been widely used in medical research to study longitudinal change, the consistency of measurement (assessment), and to establish bioequivalence. The methods presented here are examples of each one of these instances. Therefore, these methods should be adaptable to research in healthcare simulation. Specifically, three methods are discussed: hierarchical linear models, generalizability theory, and equivalence testing.

Before moving forward, a disclaimer is needed. The statistical methods here are advanced and are being presented within the context of a general model (i.e., LMM). To keep the discussion concise, some statistical notation and equations are used. However, the models are presented in their simplest forms through examples. Therefore, the general concepts are accessible to all academics and researchers.

D. Nestel et al. (eds.), *Healthcare Simulation Research*, https://doi.org/10.1007/978-3-030-26837-4_30

Hierarchical Linear Models

An intuitive form of the LMM is the hierarchical linear model (HLM) [2, 3]. A key distinction of HLMs is that they are specifically formulated to account for a clustered (nested) data structure. The simplest clustered data structure is when the units of analysis are nested within a cluster. Such clustered structures can occur in organizations and individual change. An organizational example is when students (units) are nested within medical schools (clusters). An example of individual change is when repeated measures are made of each individual in a study. In this example, the repeated measures (units) are nested within individuals (clusters). These two separate clustered structures can also be combined. Suppose an obesity study is being conducted at multiple clinics in the country in which participants are weighed multiple times over the duration of the study. In this situation, the repeated weight measures are nested within each participant, and the participants are nested within the clinics.

Consider data in which airway management skills (AMS) are measured over time (a_t; 4 occasions, 2 months apart) for paramedic trainees that received one of two training methods: modified simulation (ms; $n_{ms} = 16$) or standard simulation (ss; $n_{ss} = 11$). In such a situation, a split-plot ANOVA is the standard way to analyze AMS with time as the within-subjects factor, and training method as the between-subjects factor. Table 30.1 presents the results indicating significant method and time main effects. These effects would typically be investigated with post hoc tests. However, an alternative is to approach the whole analysis through HLM.

HLM specifies models by breaking them up into levels that account for the clustered structure of the data. A level is added for every clustering in the data. For this reason, HLM is also commonly referred to as multilevel modeling. The current example constitutes a two-level model in which time (units) is nested within paramedic trainees (clusters) and can be captured through a random-coefficient regression model (or random coefficient model).

The level-1 model can model time linearly through regression for each trainee and takes the following form:

$$y_{ti} = \pi_{0i} + \pi_{1i} a_{ti} + e_{ti}. \tag{30.1}$$

The model has an intercept (π_{0i}), slope (π_{1i}), and residual (e_{ti}) for each trainee. The intercept is the AMS at start for each trainee. The slope is the 2-month AMS change for each trainee; i.e., how much does AMS change every 2 months. The residual is assumed be independently normally distributed with constant variance σ^2. The level-1 model essentially models where the trainees started and how much they changed over the course of the study.

There are two things to point out about the level-1 model. First, this is the simplest form the level-1 model can take for time. It can be expanded to include higher order terms as needed. Second, the spacing between measurements can be different for the individuals; i.e., trainees do not have to be measured exactly every two years.

The level-2 model takes the following form:

$$\pi_{0i} = \beta_{00} + \beta_{01} ms_i + r_{0i} \tag{30.2}$$

$$\pi_{1i} = \beta_{10} + \beta_{11} ms_i + r_{1i}. \tag{30.3}$$

Notice that now the intercept (π_{0i}) and slope (π_{1i}) from level-1 are each modeled through regression. The model can now be described in terms of fixed (βs) and random effects (r_{0i} and r_{1i}). The first set of fixed effects are the average AMS for the standard simulation (β_{00}) and the average distance difference for the modified simulation (β_{01}) at start. The second set of fixed effects are the average 2-month distance slope for the standard simulation (β_{10}) and the slope difference for the modified simulation (β_{11}).

The random effects are captured with r_{0i} and $r_{1i,}$ which are assumed to be normally distributed with variances τ_{00} and τ_{11}, respectively. Here, τ_{00} captures the variability in π_{0i} (i.e., how much the trainees vary in AMS at start), and τ_{11} captures the variability in π_{1i} (i.e., how much the trainees vary in their change). An additional component not explicitly shown in the models above is the covariance τ_{01} between the intercept (r_{0i}) and slope (r_{1i}) random effects. Now the relationship between where trainees start and how much they change can be estimated.

The fixed effects are presented in Table 30.2. First, there is no significant AMS difference between the modified simulation and standard simulation at start (p-value = .088). Second, AMS for the standard simulation significantly increases over the time of the study (p-value < .001).

Table 30.1 ANOVA table for Airway Management Skills (AMS)

Source	SS	df	MS	F	p-value
Between					
Method (M)	140.465	1	140.465	9.292	.005
Error	377.915	25	15.117		
Within					
Time (T)	209.437	3	69.812	35.347	<.001
M × T	13.993	3	4.664	2.362	.078
Error	148.128	75	1.975		

Table 30.2 Fixed effects of random-coefficient regression model

Fixed effects	Estimate	SE	t-test	p-value
AMS at start				
Avg. ss AMS (β_{00})	21.21	0.61	34.77	<.001
Avg. ms AMS difference (β_{01})	1.41	0.79	1.78	.088
Slope for 2-Month AMS change				
Avg. ss AMS slope (β_{10})	0.48	0.10	4.80	<.001
Avg. ms AMS slope difference (β_{11})	0.30	0.13	2.31	.026

Table 30.3 Random effects of random-coefficient regression model

Random effects	Estimate	SE	z-test	*p*-value
Residual variance (σ^2)	1.72	0.33	5.21	<.001
Intercept variance (τ_{00})	2.91	1.14	2.55	.006
Slope variance (τ_{11})	0.02	0.03	0.67	.243
Intercept-slope covariance (τ_{10})	−.01	0.15	−.07	.956

However, AMS for the modified simulation method significantly increases at a faster rate than the standard simulation (*p*-value = .026).

The random effects are presented in Table 30.3. First, the level-1 residual variance is significant, indicating that there is unexplained variance in the model. Perhaps adding a quadratic term that can model time curvilinearly at level-1 can help explain more variance and improve the fit of the model. Second, the intercept variance is significant, indicating that trainee AMS varies at start. Third, slope variance is not significant suggesting that trainees do not vary in their rate of AMS change. Lastly, the intercept-slope covariance is not significant, so there is no relationship between AMS at start and its rate of change.

In summary, the modified simulation method is more effective than the standard simulation at improving AMS. Specifically, trainee AMS improves under both the modified simulation and standard simulation over time, but improves at a faster rate under the modified simulation. In addition, trainee AMS is similar at the start of the study for both methods, and trainee AMS improved over time regardless of their AMS at the start of the study. For HLM examples see Gadde et al. [4] and Elobeid et al. [5].

Generalizability Theory

Another form of the LMM is generalizability (G) theory [6]. However, the question(s) addressed here pertain to the consistency of measurement, and hypothesis testing is of little to no interest. Measurement is an important process in any of the sciences as it is the foundation by which data are generated. This is as true for establishing the efficacy of a medical intervention as it is for simulation-based training. Measurement is a discipline itself, but all the ideas fall into one of two equally important concepts: validity and reliability (see Chap. 26). Here, the focus is on reliability as it relates to G theory. However, G theory formulates and extends the classical true score model using a LMM. Even so, classical test theory (CTT) reliability is discussed first.

The classical true score model from CTT formulates the observed score for a measurement as

$$x = \tau + u \qquad (30.4)$$

where x is the measured data point (observed score), τ is the true score, and u is random measurement error. The idea is that every time a data point is measured, it has an element of truth (τ) plus an element of error (u). The model can be used to form the following reliability index

$$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_u^2} \qquad (30.5)$$

which is the proportion of true score variance to true score variance plus measurement error variance. The ideal situation is when there is no error (i.e., $u = 0$) as then the data point is equal to truth, and reliability would be perfect (i.e., $\rho = 1$). However, this is extremely rare in behavioral/social science research. Depending on the assumptions, the reliability index can take on different forms. If the assumption of tau-equivalence (or essentially tau-equivalence) is at least satisfied, one form that the reliability index can take is coefficient (or Cronbach's) alpha [7].

Coefficient alpha is the most common reliability index reported for a measurement instrument in many fields, including medicine and nursing [8, 9]. Coefficient alpha owes its popularity to three key features [7]. First, it is computationally simple, requiring only the number of items in the measurement instrument and the corresponding covariance matrix. Second, it can be computed for continuous, ordinal, or dichotomous items. Third, it only requires a single administration of the corresponding measurement instrument. Coefficient alpha is defined as

$$\rho = \alpha_C = \frac{k}{k-1}\left(1 - \frac{\sum_i \sigma_{ii}}{\sum_i \sum_j \sigma_{ij}}\right) \qquad (30.6)$$

where k is the number of items, $\sum_i \sigma_{ii}$ is the sum of all the k item variances, and $\sum_i \sum_j \sigma_{ij}$ is the sum of all the item variances and covariances.

For example, suppose researchers are interested in how well a set of 3 emergency medicine simulation scenarios scored by 2 raters measures knowledge of emergency medicine in junior residents. In the study, a sample of 13 junior residents participate in every scenario scored by every rater. A standard way to investigate the reliability of this design is to compute coefficient alpha for the scenarios and raters. The following covariance matrices are obtained for scenarios and raters, respectively:

$$\hat{\Sigma}_S = \begin{bmatrix} 2.59 & 2.50 & 1.29 \\ 2.50 & 4.17 & 1.75 \\ 1.29 & 1.75 & 1.56 \end{bmatrix} \quad \text{and} \quad \hat{\Sigma}_R = \begin{bmatrix} 6.86 & 3.71 \\ 3.71 & 5.14 \end{bmatrix}.$$

The corresponding coefficient alphas are $\hat{\alpha}_C = 0.86$ for scenarios, and $\hat{\alpha}_C = 0.76$ for raters. The issue here is that scenarios and raters interacted with one another as part of one design (or measurement process) and the coefficient alpha for each ignores this aspect of the design; e.g., coefficient alpha for raters ignores the impact of scenario and vice versa. This highlights a limitation of the CTT reliability methods: they can only assess one form of measurement at a time. Therefore, a method that can simultaneously assess multiple forms of measurement is required. This is precisely what G theory does.

Before moving forward, some G theory terminology must be briefly presented. In G theory anything that is used to measure is considered a source of measurement error and called a facet. The variance associated with the facets and anything they interact with is considered error variance. On the other hand, what is being measured is called the object of measurement, and the associated variance is the universe score variance (i.e., G theory's version of true score variance). In the current example, scenarios (s; $n_s = 3$) and raters (r; $n_r = 2$) are facets, and junior residents (p; $n_p = 13$) are the objects of measurement. Lastly, G theory breaks the entire analysis into two pieces: a generalizability (G) and decision (D) study. In the G study, researcher(s) obtain estimates of all the relevant variance for the measurement process. The D study is where researcher(s) obtain reliability estimates for the measurement process.

In a G study, the variability in the measurement process is captured by reformulating the classical true score model as a LMM with each facet and objects of measurement as terms in the model. As such, G theory has the same modeling flexibility as a LMM in that it can have any combination of fixed and random effects. Continuing with the current example, every junior resident participated in every scenario scored by every rater. In G theory, this constitutes a completely crossed design ($p \times s \times r$) that can be captured with the following model

$$x = \mu + p + s + r + ps + pr + sr + u \qquad (30.7)$$

where x is the observed score, μ the grand mean, p are the junior residents, s are the scenarios, r are the raters, and u is the error (residual). Although this is a LMM with all random effects, the main interest in G theory is the variability via the variance component (VC) associated with each of the terms (or sources) and their corresponding interactions (e.g., $p \times s$). The variance sources can be presented through an ANOVA table.

Table 30.4 is the ANOVA table for the knowledge of emergency medicine example. There are a few important differences to note from traditional ANOVA. First, there are no F-tests and accompanying p-values because these are not of interest in G theory. Second, what is considered the sample size in traditional ANOVA methodology is now an important source in the model (p) as it captures true differences in knowledge, skill, etc. (i.e., true score variance). Third, there is no variance for the highest order interaction ($p \times s \times r$) because there are no *df* to estimate it. This is because the sample size (i.e., p) is now a term in the model (see previous second point). Thus, the variance for the highest order interaction and error are confounded and cannot be disentangled, and together they are referred to as the residual.

Table 30.4 ANOVA table for two-facet $p \times s \times r$ design

Source	SS	df	MS	VC $\left(\hat{\sigma}^2\right)$	% of total variance
Junior residents (p)	38.82	12	3.24	0.38	0.36
Scenario (s)	1.56	2	0.78	0.01	0.01
Rater (r)	5.65	1	5.65	0.13	0.12
$p \times s$	11.10	24	0.46	0.10	0.10
$p \times r$	9.18	12	0.77	0.17	0.16
$s \times r$	0.54	2	0.27	0.00	0.00
Residual (u)	6.13	24	0.26	0.26	0.25

Note. All facets are random, $u = (p \times s \times r)$ + error

The ANOVA table is the G study and first piece of a G theory analysis. The last column in Table 30.4 contains the relative percentage of each VC. As such, the largest VC (0.36) is for junior residents (p) indicating that the most variance is universe score variance (i.e., true score variance). In terms of error, the second largest VCs are for raters (r) and its interaction with junior residents ($p \times r$). This indicates that raters are not as consistent as scenarios and are vary in their scoring of junior residents; i.e., the raters are scoring the junior residents differently. However, the largest error VC is for the residual (0.25), suggesting that something not accounted for by the G study design is impacting the measurement process. Once the VCs are estimated, they can then be used to estimate G theory reliability analogs [6].

In a D study, G theory offers two reliability analogs. The first index is the generalizability coefficient defined as

$$E\rho^2 = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\delta^2} \qquad (30.8)$$

where σ_τ^2 is universe score variance and σ_δ^2 is relative error variance defined as

$$\sigma_\delta^2 = \frac{\sigma_{ps}^2}{n_s} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{psr,e}^2}{n_s n_r}. \qquad (30.9)$$

The second is the index of dependability defined as

$$\Phi = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\Delta^2} \qquad (30.10)$$

where σ_Δ^2 is the absolute error variance defined as

$$\sigma_\Delta^2 = \frac{\sigma_s^2}{n_s} + \frac{\sigma_r^2}{n_r} + \frac{\sigma_{ps}^2}{n_s} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{sr}^2}{n_s n_r} + \frac{\sigma_{psr,e}^2}{n_s n_r} \qquad (30.11)$$

Continuing with the current example using the estimated VCs, then $\hat{\sigma}_\delta^2 = 0.16$ and $\hat{\sigma}_\Delta^2 = 0.23$. Using these quantities with $\hat{\sigma}_\tau^2 = \hat{\sigma}_p^2 = 0.38$, then $E\hat{\rho}^2 = 0.70$ and $\hat{\Phi} = 0.62$.

In the behavioral/social sciences, a typical criterion for adequate reliability is .70 or higher [10]. Because the current measurement process with 3 items and 2 raters is at or below .70, its reliability is questionable. While there does not appear to be such a criterion for G theory reliability indices, the .70 criterion does give a good benchmark from which to start. Of course, the criterion is contingent on the discipline and purpose of the measurement process. Even though both G theory reliability indices were demonstrated, each has a specific use when making decisions about the objects of measurement [6]. The generalizability coefficient is only appropriate when making relative decisions, and the index of dependability when making absolute decisions. Relative decisions are based on comparing the objects of measurement to one another; i.e., how a person compares with other people. Absolute decisions are based on comparing objects of measurement to the pre-established criterion for what is being measured; i.e., does a person meet a certain skill level. For G theory examples see McBride et al. [11] and Nadkarni et al. [12].

Equivalence Testing

The last method discussed is equivalence testing which originated in pharmacokinetics for establishing practical similarity (bioequivalence) between groups [13]. A typical situation in pharmacokinetics is a pharmaceutical company wanting to determine if a generic drug is as effective as the current drug. As such, establishing statistical equivalence is growing in popularity across the sciences outside of pharmacokinetics in a variety of settings.

Equivalence testing is the simplest form the LMM can take but is probably the most difficult to grasp. This is because the same model and corresponding results are used to test seemingly opposing hypotheses than traditionally done in null significance hypothesis testing (NSHT). NSHT and equivalence testing are flip sides of the same coin. Each method sets up two opposing hypotheses: the null (H_0) and alternative (H_A) hypothesis. Both methods assume H_0 to be true unless data provide sufficient evidence to reject it. The difference between the methods lies in how these hypotheses are stated. Consider the situation involving two means. In a typical NSHT scenario, H_0 states that the mean difference is equal to zero and H_A that the mean difference is not equal to zero. On the other hand, equivalence testing states H_0 as the mean difference surpassing or being equal to Δ and H_A as the mean difference being within Δ, where Δ is a content specific value chosen by the researcher(s) using literature, prior knowledge, or expertise.

To illustrate, consider a study in which medical students are trained in patient-centered communication through an online interactive tool. The students (T) and professionals in the field (C) are then asked to view a 6-min video of a clinical scenario and assess the care providers' communication behavior. The following assessment estimates are obtained: $n_T = 30$, $\hat{\mu}_T = 25.7$, $\hat{\sigma}_T = 1.8$; $n_C = 32$, $\hat{\mu}_C = 24.6$, $\hat{\sigma}_C = 2.1$. The standard way to compare the means from the two conditions is an independent-samples t-test which is the simplest form of the LMM with only one fixed effect. In traditional NSHT, the idea is to test for a difference between the two conditions. Here, the corresponding hypotheses can take on the following forms

$$\begin{aligned} H_0 &: \mu_T - \mu_C = 0 \\ H_A &: \mu_T - \mu_C \neq 0 \end{aligned}. \qquad (30.12)$$

The hypotheses above set up a two-sided test with $df = 60$ that gives a critical value of $t_{crit} = \pm 2.00$ using $\alpha = .05$. The corresponding independent-samples t-test is $t = 2.21$.

A two-sided test provides two options for proceeding with hypothesis testing. The first option uses the following criteria: if a test statistic surpasses the critical value, reject H_0. For the current example, H_0 can be rejected because $t = 2.21 > t_{crit} = 2$. The second option considers the following criteria: if zero is not within the confidence interval (CI), reject H_0. For the current example, the 95% CI is [0.103, 2.097], and H_0 can be rejected because zero is not within the CI. In either case, it can be concluded that students gave more favorable assessments than the professionals.

By contrast, suppose the researcher wants to test for equivalence between students and professionals. In addition, based on prior assessment studies, the researcher specifies an assessment difference of Δ = 2.5 as not meaningful. This is a situation for equivalence testing via two-one-sided t-tests (TOST) [14, 15]. Here, the corresponding hypotheses take on the following forms

$$\begin{aligned} H_0 &: \left|\mu_T - \mu_C\right| \geq \Delta \\ H_A &: \left|\mu_T - \mu_C\right| < \Delta \end{aligned}. \qquad (30.13)$$

TOST sets up two composite hypotheses based on H_0. The lower H_0 takes the following form

$$H_{0L} : \hat{\mu}_T - \hat{\mu}_C \leq -\Delta \qquad (30.14)$$

with corresponding independent-samples t-test $t_L = 7.22$. The upper H_0 takes the following form

$$H_{0U} : \hat{\mu}_T - \hat{\mu}_C \geq \Delta \qquad (30.15)$$

with corresponding independent-samples t-test $t_U = -2.81$. These are two one-sided t-tests that are corrected for Type I error by dividing α by the number of tests (i.e., Bonferroni procedure). With $df = 60$, the critical values are $t_{crit} = \pm 2.00$ using $\alpha/2 = .05/2 = .025$.

The TOST procedure also provides two options for proceeding with hypothesis testing. The first option considers

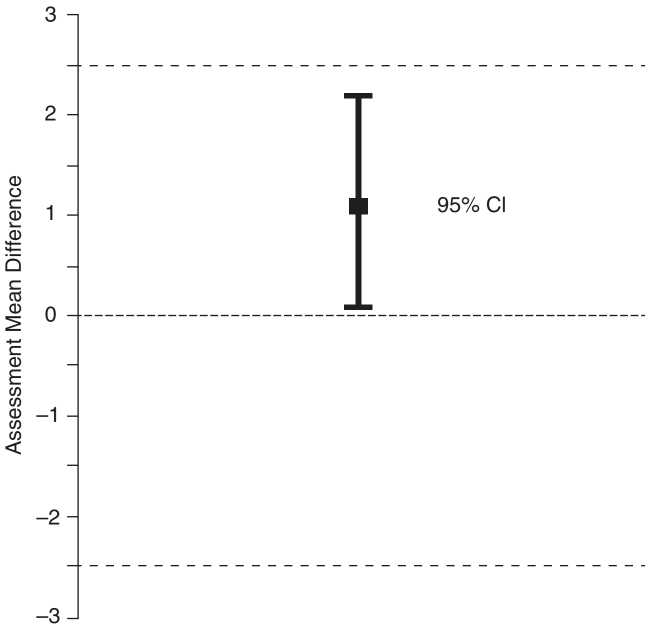

Fig. 30.1 Assessment mean difference 95% confidence interval (CI) with equivalence bounds (±2.5)

the following criteria: if the lower test statistic (t_L) is greater than a positive critical value and the upper test statistic (t_U) is less than a negative critical value, reject H_0. For the current example, H_0 can be rejected because $t_L = 7.22 > t_{crit} = 2$ and $t_U = -2.81 < t_{crit} = -2$. The second option considers the following criteria: if the CI is within ±Δ, reject H_0. For the current example, the 95% CI is [0.103, 2.097], and H_0 can be rejected because the CI is within ±2.5. In either case, the assessment of the students is practically equivalent to the professionals. The idea behind equivalence testing can be succinctly presented in a graph. Figure 30.1 presents the 95% CI along with the equivalence bounds (±Δ) where it is clear that the CI is within the equivalence bounds. For an equivalence testing example see Anderson-Montoya et al. [16].

Conclusion

HLM, G theory, and equivalence testing were briefly presented. However, this brief presentation does not do any of these methods justice and the reader is referred to the corresponding references for further details. Additionally, the methods were presented in the context of a LMM to show that, even though these methods answer different questions, they share a general statistical framework. As such, any of the standard statistical packages (e.g., SAS, SPSS, R, etc.) through their LMM routines can run any of the models. However, the packages only compute the G theory VCs but not the corresponding reliability indices. The VCs can be used to hand-compute or use software (e.g., MS Excel) to get the required reliability estimates ($E\rho^2$, Φ). To get all the G theory estimates, then GENOVA [6] or EduG [17] can be used. For equivalence testing, a simple t-test routine from the packages can be used.

In conclusion, three statistical methods commonly used in medical research were presented. Each method answers different research questions and hence have different applications. Therefore, each method was presented through an application. In each application, the advantage of the method is demonstrated by contrasting it with the traditional method of analysis. Through this process it was made clear that HLM and G theory offer more flexibility and provide a richer analysis than traditional ANOVA and coefficient alpha, respectively. By contrast, equivalence testing is not necessarily richer than a NSHT, but it does demonstrate how the same results can be used to answer seemingly opposing hypotheses. Although the hypotheses are opposing, both have their place in research. It is hoped the presentation here has sparked the curiosity of healthcare simulation researchers and given ideas as to how to adapt the methods in their research.

References

1. Muller KE, Stewart PW. Linear model theory: univariate, multivariate, and mixed models. Hoboken: Wiley-Interscience; 2006. xiv, p. 410.
2. Raudenbush SW, Bryk AS. Hierarchical linear models: applications and data analysis methods. 2nd ed. Thousand Oaks: Sage Publications; 2002. xxiv, p. 485.
3. Snijders TAB, Bosker RJ. Multilevel analysis: an introduction to basic and advanced multilevel modeling. 2nd ed. Los Angeles: Sage; 2012. xi, p. 354
4. Gadde KM, Franciscy DM, Wagner HR, Krishnan KRR. Zonisamide for weight loss in obese adults – a randomized controlled trial. Jama-J Am Med Assoc. 2003;289(14):1820–5.
5. Elobeid MA, Padilla MA, McVie T, Thomas O, Brock DW, Musser B, et al. Missing data in randomized clinical trials for weight loss: scope of the problem, state of the field, and performance of statistical methods. PLoS One. 2009;4(8):e6624.
6. Brennan RL. Generalizability theory. New York: Springer; 2001. xx, p. 538
7. Padilla MA, Divers J, Newton M. Coefficient Alpha bootstrap confidence interval under nonnormality. Appl Psychol Meas. 2012;36(5):331–48.
8. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78(1):98–104.
9. Hogan TP, Benjamin A, Brezinski KL. Reliability methods: a note on the frequency of use of various types. Educ Psychol Meas. 2000;60(4):523–31.

10. Peterson RA. A meta-analysis of cronbach's coefficient alpha. J Consum Res. 1994;21(2):381–91.
11. McBride ME, Waldrop WB, Fehr JJ, Boulet JR, Murray DJ. Simulation in pediatrics: the reliability and validity of a multiscenario assessment. Pediatrics. 2011;128(2):335–43.
12. Nadkarni LD, Roskind CG, Auerbach MA, Calhoun AW, Adler MD, Kessler DO. The development and validation of a concise instrument for formative assessment of team leader performance during simulated pediatric resuscitations. Simul Healthc. 2018;13(2):77–82.
13. Hauck WW, Anderson S. A new statistical procedure for testing equivalence in two group comparative bioavailability trials. J Pharmacokinet Biopharm. 1984;12(1):83–91.
14. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987;15(6):657–80.
15. Wellek S. Testing statistical hypotheses of equivalence and noninferiority. 2nd ed. Boca Raton: CRC Press; 2010. xvi, p. 415.
16. Anderson-Montoya BL, Scerbo MW, Ramirez DE, Hubbard TW. Running memory for clinical handoffs: a look at active and passive processing. Hum Factors. 2017;59(3):393–406.
17. Cardinet J, Johnson S, Pini G. Applying generalizability theory using EduG. New York: Routledge; 2010. xviii, p. 215.